JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 8 (2020), Issue 7 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Original Papers

Viewpoint

Review

Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review

Avishek Choudhury[1], MSc; Onur Asan[1], PhD

School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, United States

Corresponding Author:
Onur Asan, PhD
School of Systems and Enterprises
Stevens Institute of Technology
1 Castle Point Terrace
Hoboken, NJ, 07030
United States
Phone: 1 2012168901 ext 2012168901
Email: oasan@stevens.edu

Abstract

Background: Artificial intelligence (AI) provides opportunities to identify the health risks of patients and thus influence patient safety outcomes.

Objective: The purpose of this systematic literature review was to identify and analyze quantitative studies utilizing or integrating AI to address and report clinical-level patient safety outcomes.

Methods: We restricted our search to the PubMed, PubMed Central, and Web of Science databases to retrieve research articles published in English between January 2009 and August 2019. We focused on quantitative studies that reported positive, negative, or intermediate changes in patient safety outcomes using AI apps, specifically those based on machine-learning algorithms and natural language processing. Quantitative studies reporting only AI performance but not its influence on patient safety outcomes were excluded from further review.

Results: We identified 53 eligible studies, which were summarized concerning their patient safety subcategories, the most frequently used AI, and reported performance metrics. Recognized safety subcategories were clinical alarms (n=9; mainly based on decision tree models), clinical reports (n=21; based on support vector machine models), and drug safety (n=23; mainly based on decision tree models). Analysis of these 53 studies also identified two essential findings: (1) the lack of a standardized benchmark and (2) heterogeneity in AI reporting.

Conclusions: This systematic review indicates that AI-enabled decision support systems, when implemented correctly, can aid in enhancing patient safety by improving error detection, patient stratification, and drug management. Future work is still needed for robust validation of these systems in prospective and real-world clinical environments to understand how well AI can predict safety outcomes in health care settings.

(JMIR Med Inform 2020;8(7):e18599) doi:10.2196/18599

KEYWORDS

artificial intelligence; patient safety; drug safety; clinical error; report analysis; natural language processing; drug; review

Introduction

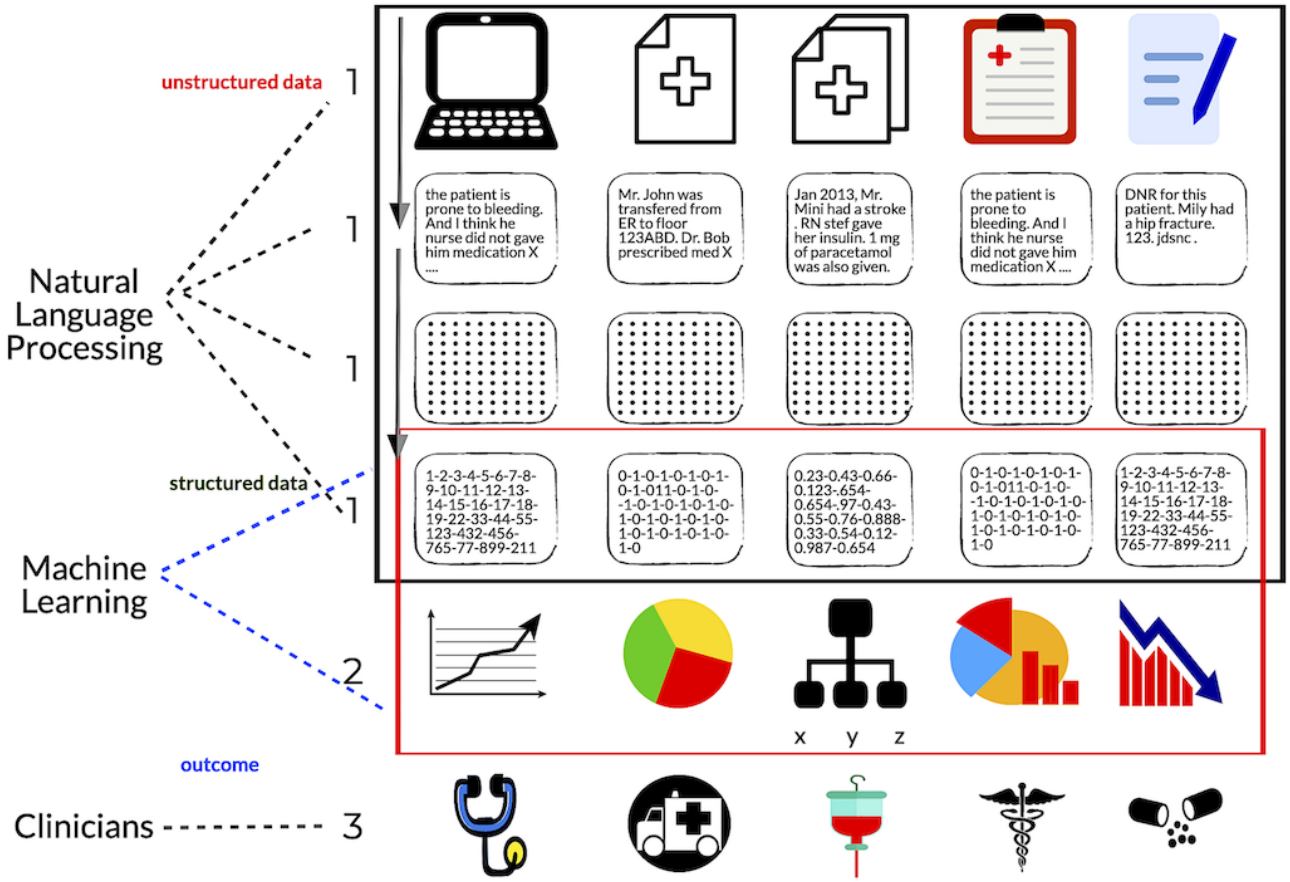
Patient safety is defined as the absence of preventable harm to a patient and minimization of the risk of harm associated with the health care process [1,2]. Every part of the care-giving process involves a certain degree of inherent risk. Since resolution WHA55.18 on “Quality of Care: Patient Safety” at the 55th World Health Assembly was proposed in 2002, there has been increasing attention paid to patient safety concerns and adverse events in health care settings [3]. Despite the safety initiatives and investments made by federal and local governments, private agencies, and concerned institutions, studies continue to report unfavorable patient safety outcomes [4,5].

The integration of artificial intelligence (AI) into the health care system is not only changing dynamics such as the role of health care providers but is also creating new potential to improve patient safety outcomes [6] and the quality of care [7]. The term

AI can be broadly defined as a computer program that is capable of making intelligent decisions [8]. The operational definition of AI we adopt in this review is the ability of a computer or health care device to analyze extensive health care data, reveal hidden knowledge, identify risks, and enhance communication [9]. In this regard, AI encompasses machine learning and natural language processing. Machine learning enables computers to utilize labeled (supervised learning) or unlabeled (unsupervised learning) data to identify latent information or make predictions about the data without explicit programming [9]. Among different types of AI, machine learning and natural language processing specifically have societal impacts in the health care domain [10] and are also frequently used in the health care field [9-12].

The third category within machine learning is known as reinforcement learning, in which an algorithm attempts to accomplish a task while learning from its successes and failures [9]. Machine learning also encompasses artificial neural networks or deep learning [13]. Natural language processing focuses on building a computer's ability to understand human language and consecutively transform text to machine-readable structured data, which can then be analyzed by machine-learning techniques [14]. In the literature, the boundary defining natural language processing and machine learning is not clearly defined. However, as illustrated in Figure 1, studies in the field of health care have been using natural language processing in conjunction with machine-learning algorithms [15].

Figure 1. Schematic illustration of how natural language processing converts unstructured text to machine-readable structured data, which can then be analyzed by machine-learning algorithms.



AI has potential to assist clinicians in making better diagnoses [16-18], and has contributed to the fields of drug development [19-21], personalized medicine, and patient care monitoring [14,22-24]. AI has also been embedded in electronic health record (EHR) systems to identify, assess, and mitigate threats to patient safety [25]. However, with the deployment of AI in health care, several risks and challenges can emerge at an individual level (eg, awareness, education, trust), macrolevel (eg, regulation and policies, risk of injuries due to AI errors), and technical level (eg, usability, performance, data privacy and security).

The measure of AI accuracy does not necessarily indicate clinical efficiency [26]. Another common measure, the area under the receiver operating characteristic curve (AUROC), is also not necessarily the best metric for clinical applicability [27]. Such AI metrics might not be easily understood by clinicians or might not be clinically meaningful [28]. Moreover, AI models have been evaluated using a variety of parameters and report different measure(s) such as the *F1* score, accuracy, and false-positive rate, which are indicative of different aspects of AI's analytical performance. Understanding the functioning of complex AI requires technical knowledge that is not common

among clinicians. Moreover, clinicians do not necessarily have the training to identify underlying glitches of the AI, such as data bias, overfitting, or other software errors that might result in misleading outcomes. Such flaws in AI can result in incorrect medication dosage and poor treatment [29-33].

Furthermore, a system error in a widely used AI might lead to mass patient injuries compared to a limited number of patient injuries due to a provider's error [34]. Additionally, there have been instances where traditional analytical methods outperformed machine-learning techniques [9]. Owing to the wide range of effectiveness of AI, it is crucial to understand both the promising and deterring impacts of AI on patient safety outcomes [35].

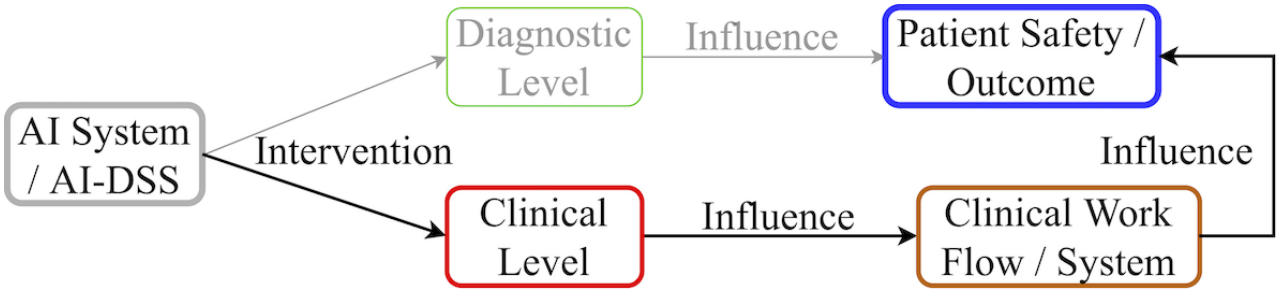
AI in the health care system can assist at both the "clinical" and "diagnostic" levels [36]. AI provides a powerful tool that can be implemented within the health care domain to reveal subtle patterns in data, and these patterns can then be interpreted by clinicians to identify new clinical and health-related issues [9]. Recent studies and reviews have primarily focused on the performance of AI at the diagnostic level, such as for disease identification [37-42], and the application of AI robotics in surgery and disease management [43-46]. Other studies have also implemented AI technologies to assist at the clinical level, including assessing fall risks [47] and medication errors [48,49]. However, many of these studies are centered around AI development and performance and there is a notable lack of studies reviewing the role and impact of AI used at the clinical level on patient safety outcomes.

Many studies have reported high accuracy of AI in health care. However, its actual influence (negative or positive) can only be realized when it is integrated into clinical settings or interpreted and used by care providers [50]. Therefore, in our view, patient safety and AI performance might not necessarily complement each other. AI in health care depends on data sources such as EHR systems, sensor data, and patient-reported data. EHR systems may contain more severe cases for specific patient populations. Certain patient populations may have more ailments or may be seen at multiple institutions. Certain subgroups of patients with rare diseases may not exist in sufficient numbers for a predictive analytic algorithm. Thus, clinical data retrieved from EHRs might be prone to biases [9,50]. Owing to these potential biases, AI accuracy might be misleading [51] when trained on a small subgroup or small sample size of patients with rare ailments.

Furthermore, patients with limited access to health care may receive fewer diagnostic tests and medications and may have insufficient health information in the EHR to trigger an early intervention [52]. In addition, institutions record patient information differently; as a result, if AI models trained at one institution are implemented to analyze data at another institution, this may result in errors [52]. For instance, machine-learning algorithms developed at a university hospital to predict patient-reported outcome measures, which tend to be documented by patients who have high education as well as high income, may not be applicable when implemented at a community hospital that primarily serves underrepresented patient groups with low income.

A review [53] conducted in 2017 reported that only about 54% of studies that developed prediction models based on EHRs accounted for missing data. Recent studies and reviews have been primarily focusing on the performance and influence of AI systems at a diagnostic level, such as for disease identification [37-42], and the influence of AI robotics in surgery and disease management [43-46]; however, there is a lack of studies reviewing and reporting the impact of AI used at the clinical level on patient safety outcomes, as well as characteristics of the AI algorithms used. Thus, it is essential to study how AI has been shown to influence patient safety outcomes at the clinical level, along with reported AI performance in the literature. In this systematic review, we address this gap by exploring the studies that utilized AI algorithms as defined in this review to address and report changes in patient safety outcomes at the clinical level (Figure 2).

Figure 2. Artificial intelligence (AI) route to patient safety via "Clinical" and "Diagnostic" level interventions. DSS: decision support system.



Methods

Protocol Registration

This systematic review is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [54]. We followed the PRISMA Checklist (see Multimedia Appendix 1). Our protocol [55] was registered with the Open Science Framework on September 15, 2019.

Information Sources

We searched for peer-reviewed publications in the PubMed, PubMed Central, and Web of Science databases from January 2009 to August 2019 to identify articles within the scope and eligibility criteria of this systematic literature review.

Search Strategy

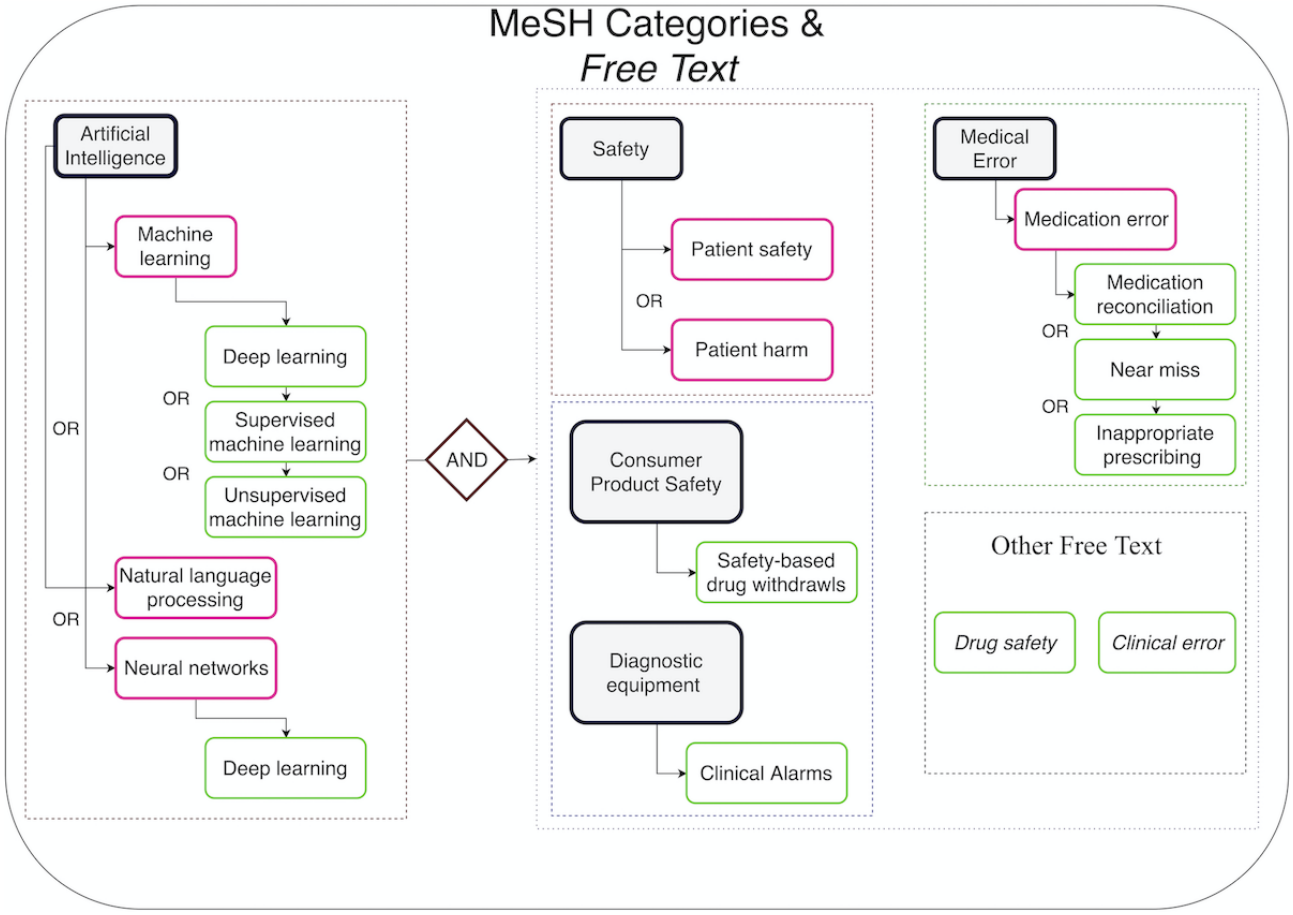
We followed a systematic approach of creating all search terms to capture all related and eligible papers in the searched

databases. Keywords used in the search were initially determined by a preliminary review of the literature and then modified based on feedback from content experts as well as our institution's librarian.

We then refined the search strategy in collaboration with the librarian to ensure that all clinical-level patient safety-related papers (as shown in Figure 2) were covered in our review and determined the Medical Subject Headings (MeSH) terms. We grouped the query keywords, which were derived from MeSH terms and combined through Boolean (AND/OR) operators to identify all relevant studies that matched with our scope and inclusion criteria.

The keywords consisted of MeSH terms such as "safety [MeSH]" and "artificial intelligence [MeSH]," in combination with narrower MeSH terms (subheadings/related words/phrases) and free text for "artificial intelligence" and "safety." We also included broader key terms to encompass all latent risk factors affecting patient safety. The final search keywords (Figure 3) described below were used to explore all databases.

Figure 3. Medical Subject Heading (MeSH) terms and free text used in the systematic literature review.



MeSH terms are organized in a tree-like hierarchy, with more specific (narrower) terms arranged beneath broader terms. By default, PubMed includes all of the narrow items in the search in a strategy known as "exploding" the MeSH term [56]. Moreover, the inclusion of MeSH terms optimizes the search strategy [56]. Therefore, the final search query for PubMed was as follows: ("patient safety" OR "safety" [MeSH] OR "drug safety" OR "safety-based Drug withdraws" [MeSH] OR "medication error" OR "Medication Error" [MeSH] OR "medication reconciliation" OR "near miss" OR "inappropriate prescribing" OR "clinical error" OR "Clinical alarms" [MeSH]) AND ("Machine learning" [MeSH] OR "Machine learning" OR "Deep learning" [MeSH] OR "Deep learning" OR "natural language processing" [MeSH] OR "natural language processing").

Inclusion and Exclusion Criteria

This study focused on peer-reviewed publications satisfying the following two primary conditions: (1) implementation of machine-learning or natural language processing techniques to address patient safety concerns, and (2) discussing or reporting the impact or changes in clinical-level patient safety outcomes. Any papers that failed to satisfy both conditions were excluded from this review. For instance, studies only focusing on developing or evaluating machine-learning models that did not report or discuss changes or impact on clinical-level patient safety outcomes were excluded, as well as studies that used AI beyond our scopes, such as robotics or computer vision. Secondary research such as reviews, commentaries, and conceptual articles was excluded from this study. The search was restricted to papers published in English between January 2009 and August 2019.

Study Selection and Quality Assurance

The two authors together reviewed all of the retrieved publications for eligibility. We first screened the publications by studying the titles and abstracts and removed duplications. We then read the full text for the remaining papers and finalized the selection. To minimize any selection bias, all discrepancies were resolved by discussion requiring consensus from both reviewers and the librarian. Before finalizing the list of papers, we consulted our results and searched keywords with the librarian to ensure that no relevant articles were missed.

A data abstraction form was used to record standardized information from each paper as follows: authors, aims, objectives of the study, methods, and findings. Using this form, we categorized each article based on the type of AI algorithm as well as clinical-level patient safety outcomes reported.

Results

Study Selection

Figure 4 illustrates the flowchart of the selection process of the articles included in this systematic literature review. The initial search using a set of queries returned 272 publications in PubMed, 1976 publications in PubMed Central, and 248 publications in Web of Science for a total of 2496 articles. We used EndNote X9.3.2 to manage the filtering and duplication removal process. As a first step, we removed duplicates (n=101), all review/opinion/perspective papers (n=120), and posters or short abstracts (n=127). The two authors then applied a second filtering step by reading abstracts and titles (n=2148). The screening process followed the inclusion and exclusion criteria explained above, resulting in 80 papers eligible for a full-text review. The authors then removed 27 more articles based on the full-text review. Hence, the final number of studies included in the systematic review was 53, with consensus from both authors.

Figure 4. Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) flow chart illustrating the process of selecting eligible publications for inclusion in the systematic review. WoS: Web of Science.

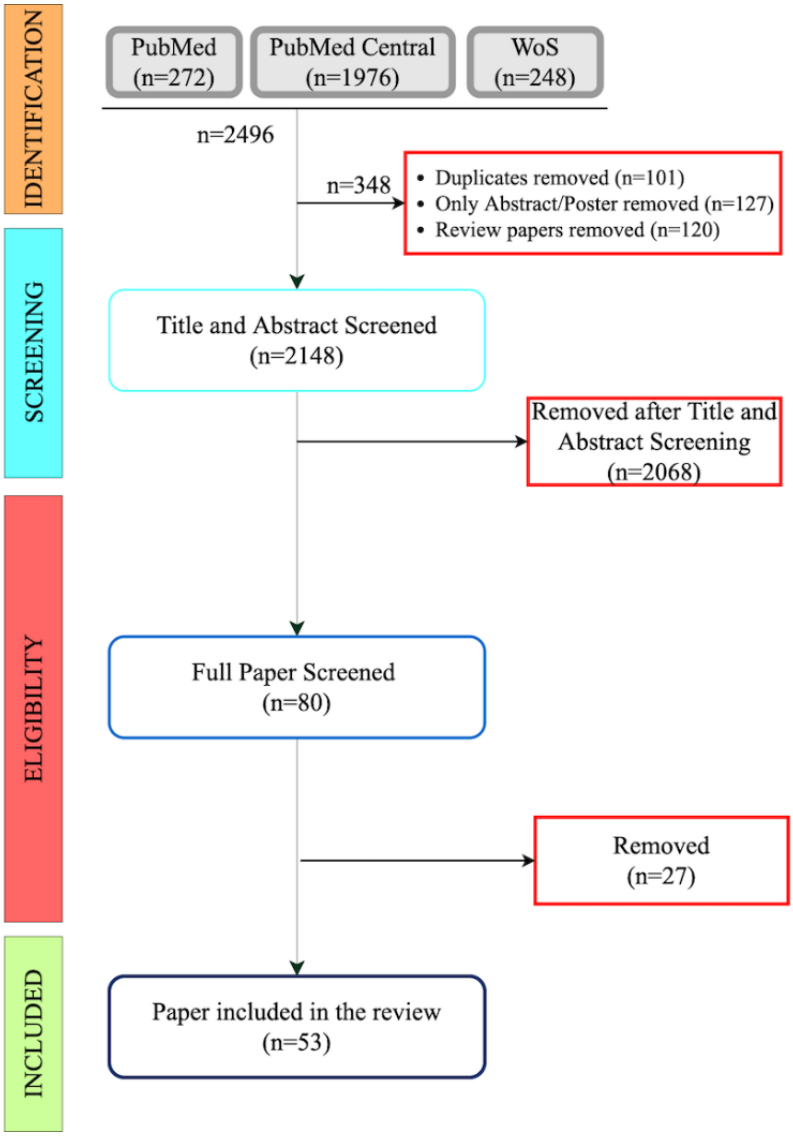


Table 1 outlines all characteristics of the final selected studies (n=53), including the objective of the study and AI methods used, as well as classification of all articles by latent risk factors of patient safety according to (a) Clinical Alarms/Alerts, (b) Clinical Reports, and (c) Adverse Drug Event/Drug Safety. Table 1 also reports the findings obtained regarding changes in patient safety outcomes.

The studies mostly reported positive changes in patient safety outcomes, and in most cases improved or outperformed traditional methods. For instance, AI was successful in minimizing false alarms in several studies and also improved real-time safety reporting systems (Table 1). AI was also able to extract useful information from clinical reports. For example, AI helped in classifying patients based on their ailments and severity, identified common incidents such as fall risks, delivery delays, hospital information technology errors, bleeding complications, and others that pose risks to patient safety. AI also helped in minimizing adverse drug effects. Further, some studies reported poor outcomes of AI, in which AI's classification accuracy was lower than that of clinicians or existing standards.

Table 2 outlines the performance and accuracy measures of AI models used by the final selected studies, demonstrating the heterogeneity in AI performance measures adopted by different studies.

Table 1. Evidentiary table of 53 selected publications.

Reference	Objective	Study theme	AI[a] method	Findings (patient safety outcomes)
Chen et al [57]	To classify alerts as real or artifacts in online noninvasive vital sign data streams and minimize alarm fatigue and missed true instability	Clinical alarms/alerts	KNN[b], NB[c], LR[d], SVM[e], RF[f]	Machine-learning (ML) models could distinguish clinically relevant pulse arterial O_2 saturation, blood pressure, and respiratory rate from artifacts in an online monitoring dataset (AUC[g]>0.87)
Ansari et al [58]	To minimize false alarms in the ICU[h]	Clinical alarms/alerts	MMD[i], DT[j]	ML algorithm along with MMD was effective in suppressing false alarms
Zhang et al [59]	To minimize the rate of false critical arrhythmia alarms	Clinical alarms/alerts	SVM	SVM reduced false alarm rates. The model gave an overall true positive rate of 95% and true negative rate of 85%
Antink et al [60]	To reduce false alarms by using multimodal cardiac signals recorded from a patient monitor	Clinical alarms/alerts	BCT[k], SVM, RF, RDAC[l]	A false alarm reduction score of 65.52 was achieved; employing an alarm-specific strategy, the model performed at a true positive rate of 95% and true negative rate of 78%. False alarms for extreme tachycardia were suppressed with 100% sensitivity and specificity
Eerikäinen et al [61]	To classify true and false cardiac arrhythmia alarms	Clinical alarms or alerts	RF	Out of 5 false alarms, 4 were suppressed; 77.39% real-time model accuracy
Menard et al [62]	Develop a predictive model that enables Roche/Genentech Quality Program Leads oversight of adverse event reporting at the program, study, site, and patient level.	Clinical alarms/alerts	ML (not disclosed)	The ML method identified the sites by risk of underreporting and enabled real-time safety reporting. The proposed model had an AUC of 0.62, 0.79, and 0.92 for simulation scenarios of 25%, 50%, and 75%, respectively. This project was part of a broader effort at Roche/Genentech Product Development Quality to apply advanced analytics to augment and complement traditional clinical quality assurance approaches
Segal et al [63]	To determine the clinical usefulness of medication error alerts in a real-life inpatient setting	Clinical alarms or alerts	Probabilistic ML	85% of the alerts were clinically valid, and 80% were considered clinically useful; 43% of the alerts caused changes in subsequent medical orders. Thus, the model detected medication errors
Hu et al [64]	To detect clinical deterioration	Clinical alarms or alerts	NN[m]	NN-based model could detect health deterioration such as heart rate variability with more accuracy than one of the best-performing early warning scores (ViEWS). The positive prediction value of NN was 77.58% and the negative prediction value was 99.19%
Kwon et al [65]	To develop alarm systems that predict cardiac arrest early	Clinical alarms or alerts	RF, LR, DEWS[n], and MEWS[o]	The DEWS identified more than 50% of patients with in-hospital cardiac arrest 14 hours before the event. It allowed medical staff to have enough time to intervene. The AUC and AUPRC[p] of DEWS was 0.85 and 0.04, respectively, and outperformed MEWS with AUC and AUPC of 0.60 and 0.003, respectively; RF with AUC and AUPC of 0.78 and 0.01, respectively; and LR with AUC and AUPRC of 0.61 and 0.007, respectively. DEWS reduced the number of alarms by 82.2%, 13.5%, and 42.1% compared with the other models at the same sensitivity
Gupta and Patrick [66]	To classify clinical incidents	Clinical Report	J48[q], NB multinomial, and SVM	The selected models performed poorly in classifying incident categories (48.77% best, using J48), but performed comparatively better in classifying free text (76.49% using NB).
Wang et al [67]	To identify multiple incident types from a single report	Clinical Report	Compares binary relevance, CC[r]	Binary classifier improved identification of common incident types: falls, medications, pressure injury, aggression, documentation problem, and others. Automated identification enabled safety problems to be detected and addressed in a more timely manner

Reference	Objective	Study theme	AI[a] method	Findings (patient safety outcomes)
Zhou et al [49]	To extract information from clinical reports	Clinical Report	SVM, NB, RF, and MLP[s]	ML algorithms identified the medication event originating stages, event types, and causes, respectively. The models improved the efficiency of analyzing the medication event reports and learning from the reports in a timely manner with (SVM) *F1* of 0.792 and (RF) *F1* of 0.925
Fong et al [68]	To analyze patient safety reports	Clinical Report	NLP[t]	Pyxis Discrepancy and Pharmacy Delivery Delay were found to be the main two factors affecting patient safety. The NLP models significantly reduced the time required to analyze safety reports
El Messiry et al [69]	To analyze patient feedback	Clinical Report	NLP	Care-related complaints were influenced by money and emotion
Chondrogiannis et al [70]	To identify the meaning of abbreviations used in clinical studies	Clinical Report	NLP	Each clinical study document contained about 6.8 abbreviations. Each abbreviation can have 1.25 meanings on average. This helped in identification of acronyms
Liang and Gong [71]	To extract information from patient safety reports	Clinical Report	Multilabel classification methods	Binary relevance was the best problem transformation algorithm in the multilabeled classifiers. It provided suggestions on how to implement automated classification of patient safety reports in clinical settings
Ong et al [72]	To identify risk events in clinical incident reports	Clinical report	Text classifiers based on SVM	SVM performed well on datasets with diverse incident types (85.8%) and data with patient misidentification (96.4%). About 90% of false positives were found in “near-misses” and 70% of false negative occurred due to spelling errors
Taggart et al [73]	To identify bleeding events using in clinical notes	Clinical Report	NLP, SVM, CNN[u], and ET[v]	Rule-based NLP was better than the ML approach. NLP detected bleeding complications with 84.6% specificity, 62.7% positive predictive value, and 97.1% negative predictive value. It can thus be used for quality improvement and prevention programs
Denecke et al [74]	To minimize any loss of information during a doctor-patient conversation	Clinical Report	NLP	Electronic health platform provides an intuitive conversational user interface that patients use to connect to their therapist and self-anamnesis app. The app also allows data sharing among treating therapists
Evans et al [75]	To determine the incident type and the severity of harm outcome	Clinical Report	J48, SVM, and NB	The SVM classifier improved the identification of patient safety incidents. Incident reports containing deaths were most easily classified with an accuracy of 72.82%. The severity classifier was not accurate to replace manual scrutiny
Wang et al [76]	To identify the type and severity of patient safety incident reports	Clinical Report	CNN and SVM ensemble	CNN achieved high *F* scores (>85%) across all test datasets when identifying common incident types, including falls, medications, pressure injury, and aggression. It improved the process by 11.9% to 45.10% across different datasets
Klock et al [47]	To understand the root causes of falls and increase learning from fall reports for better prevention of patient falls.	Clinical Report	SVM, RF, and RNN[w]	The model identified high and low scoring fall reports. Most of the patient fall reports scores were between 0.3 and 0.4, indicating poor quality of reports
Li et al [77]	To stratify patient safety adverse event risk and predict safety problems of individual patients	Clinical Report	Ensemble-ML	The adverse event risk score at the 0.1 level could identify 57.2% of adverse events with 26.3% accuracy from 9.2% of the validation sample. The adverse event risk score of 0.04 could identify 85.5% of adverse events
Murff et al [78]	To identify postoperative surgical complications within a comprehensive electronic medical record	Clinical Report	NLP	NLP identified 82% of acute renal failure cases compared with 38% for patient safety indicators. Similar results were obtained for venous thromboembolism (59% vs 46%), pneumonia (64% vs 5%), sepsis (89% vs 34%), and postoperative myocardial infarction (91% vs 89%)

Reference	Objective	Study theme	AI[a] method	Findings (patient safety outcomes)
Wang et al [79]	To automate the identification of patient safety incidents in hospitals	Clinical Report	Text-based classifier: LR, SVM	For severity level, the *F* score for severity assessment code (SAC) 1 (extreme risk) was 87.3 and 64% for SAC4 (low risk) on balanced data. With stratified data, a high recall was achieved for SAC1 (82.8%-84%), but precision was poor (6.8%-11.2%). High-risk incidents (SAC2) and medium-risk incidents (SAC3) were often misclassified. Reports about falls, medications, pressure injury, aggression, and blood tests were identified with high recall and precision
Rosenbaum and Baron [80]	To detect Wrong Blood in Tube errors and mitigate patient harm	Clinical Report	LR, SVM	In contrast to the univariate analysis, the best performing multivariate delta check model (SVM) identified errors with a high degree of accuracy (0.97)
McKnight [81]	To improve the ability to extract clinical information from patient safety reports efficiently	Clinical Report	NLP	The semisupervised model categorized patient safety reports into their appropriate patient safety topic and avoided overlaps; 85% of unlabeled reports were assigned correct labels. It helped NCPS[x] analysts to develop policy and mitigation decisions
Marella et al [82]	To analyze patient safety reports describing health hazards from electronic health records	Clinical Report	Text mining based on: NB, KNN, rule induction	The NB kernel performed best, with an AUC of 0.927, accuracy of 0.855, and *F* score of 0.877. The overall proportion of cases found relevant was comparable between manually and automatically screened cases; 334 reports identified by the model as relevant were identified as not relevant, implying a false-positive rate of 13%. Manual screening identified 4 incorrect predictions, implying a false-negative rate of 29%
Ye et al [83]	To validate a real-time early warning system to predict patients at high risk of inpatient mortality during their hospital episodes	Clinical Report	RF, XGB[y], boosting SVM, LASSO[z], and KNN	The modified early warning system accurately predicted the possibility of death for the top 13.3% (34/255) of patients at least 40.8 hours before death
Fong et al [84]	To identify health information technology-related events from patient safety reports	Clinical Report	Unigram and Bigram LR, SVM	Unigram models performed better than Bigram and combined models. It identified HIT[aa]-related events trained on PSE[bb] free-text descriptions from multiple states and health care systems. The unigram LR model gave an AUC of 0.931 and an *F1* score of 0.765. LR also showed potential to maintain a faster runtime when more reports are analyzed. The final HIT model had less complexity and was more easily sharable
Simon et al [85]	To establish whether patients with type 2 diabetes can safely use PANDIT[cc] and whether its insulin dosing advice is clinically safe	Drug safety	PANDIT	27 out of 74 (36.5%) PANDIT advice differed from those provided by diabetes nurses. However, only one of these (1.4%) was considered unsafe by the panel
Song et al [86]	To predict drug-drug interactions	Drug safety	SVM	The 10 - fold crossvalidation improved the identification of drug-drug interaction with AUC>0.97, which is significantly greater than the analogously developed ML model (0.67)
Hammann et al [87]	To identify drugs that could be suspected of causing adverse reactions in the central nervous system, liver, and kidneys	Drug safety	CHAID[dd] and CART[ee]	CART exhibited high predictive accuracy of 78.94% for allergic reactions, 88.69% for renal, and 90.22% for the liver. CHAID model showed a high accuracy of 89.74% for the central nervous system
Bean et al [88]	To predict adverse drug reactions	Drug safety	LR, SVM, DT, NLP, own model	The proposed model (own model) outperformed traditional LR, SVM, DT, and predicted adverse drug reactions with an AUC of 0.92

Reference	Objective	Study theme	AI[a] method	Findings (patient safety outcomes)
Hu et al [89]	To predict the appropriateness of initial digoxin dosage and minimize drug-drug adverse interactions	Drug safety	C4.5, KNN, CART, RF, MLP, and LR	In the non drug-drug interaction group, the AUC of RF, MLP, CART, and C4.5 was 0.91, 0.81, 0.79, and 0.784, respectively; for the drug-drug interaction group, the AUC of RF, CART, MLP, and C4.5 was 0.89, 0.79, 0.77, and 0.77, respectively. DT-based approaches and MLP can determine the initial dosage of a high-alert digoxin medication, which can increase drug safety in clinical practice
Tang et al [90]	To identify adverse drug effects from unstructured hospital discharge summaries	Drug safety	NLP	A total of 33 trial sets were evaluated by the algorithm and reviewed by pharmacovigilance experts. After every 6 trial sets, drug and adverse event dictionaries were updated, and rules were modified to improve the system. The model identified adverse events with 92% precision and recall
Hu et al [91]	To predict the dosage of warfarin	Drug safety	KNN, SVR[ff], NN-BP[gg], MT[hh]	The proposed model improved warfarin dosage when compared to the baseline (mean absolute error 0.394); reduced mean absolute error by 40.04%
Hasan et al [92]	To improve medication reconciliation task	Drug safety	LR, KNN	Collaborative filtering identified the top 10 missing drugs about 40% to 50% of the time and the therapeutic missing drugs about 50% to 65% of the time
Labovitz et al [93]	To evaluate the use of a mobile AI platform on medication adherence in stroke patients on anticoagulation therapy	Drug safety	Cell phone–based AI platform	Mean (SD) cumulative adherence based on the AI platform was 90.5% (7.5%). Plasma drug concentration levels indicated that adherence was 100% (15/15) and 50% (6/12) in the intervention and control groups, respectively
Long et al [94]	To improve the reconciliation method	Drug safety	iPad-based software tool with an AI algorithm	All patients completed the task. The software improved reconciliation; all patients identified at least one error in their electronic medical record medication list; 8 of 10 patients reported that they would use the device in the future. The entire team (clinical and patients) liked the device and preferred to use it in the future
Reddy et al [95]	To assess proof of concept, safety, and feasibility of ABC4D[ii] in a free-living environment over 6 weeks	Drug safety	ABC4D	ABC4D was safe for use as an insulin bolus dosing system. A trend suggesting a reduction in postprandial hypoglycemia was observed. The median (IQR) number of postprandial hypoglycemia episodes within 6 h after the meal was 4.5 (2.0-8.2) in week 1 versus 2.0 (0.5-6.5) in week 6 (*P*=.10). No episodes of severe hypoglycemia occurred during the study
Schiff et al [96]	To evaluate the performance and clinical usefulness of medication error alerts generated by an alerting system	Drug safety	MedAware, probabilistic ML	75% of the chart-reviewed alerts generated by MedAware were valid from which medication errors were identified. Of these valid alerts, 75.0% were clinically useful in flagging potential medication errors.
Li et al [97]	To develop a computerized algorithm for medication discrepancy detection and assess its performance on real-world medication reconciliation data	Drug safety	Hybrid system consisting of ML algorithms and NLP	The hybrid algorithm yielded precision (P) of 95.0%, recall (R) of 91.6%, and *F* value of 93.3% on medication entity identification, and *P*=98.7%, R=99.4%, and *F*=99.1% on attribute linkage. The combination of the hybrid system and medication matching system gave *P*=92.4%, R=90.7%, and *F*=91.5%, and *P*=71.5%, R= 65.2%, and *F*=68.2% on classifying the matched and the discrepant medications, respectively
Carrell et al [98]	To identify evidence of problem opioid use in electronic health records	Drug safety	NLP	The NLP-assisted manual review identified an additional 728 (3.1%) patients with evidence of clinically diagnosed problem opioid use in clinical notes.
Tinoco et al [99]	To evaluate the source of information affecting different adverse events	Drug safety	CSS[ij] (ML)	CSS detected more hospital-associated infections than manual chart review (92% vs 34%); CSS missed events that were not stored in a coded format

Reference	Objective	Study theme	AI[a] method	Findings (patient safety outcomes)
Onay et al [100]	To classify approved drugs from withdrawn drugs and thus reduce adverse drug effects	Drug safety	SVM, Boosted and Bagged trees (Ensemble)	The Gaussian SVM model yielded 78% prediction accuracy for the drug dataset, including all diseases. The ensemble of bagged tree and linear SVM models involved 89% of the accuracies for psycholeptics and psychoanalytic drugs
Cai et al [101]	To discover drug-drug interactions from the Food and Drug Administration's adverse event reporting system and thus prevent patient harm	Drug safety	Causal Association Rule Discovery (CARD)	CARD demonstrated higher accuracy in identifying known drug interactions compared to the traditional method (20% vs 10%); CARD yielded a lower number of drug combinations that are unknown to interact (50% for CARD vs 79% for association rule mining).
Dandala et al [102]	To extract adverse drug events from clinical narratives and automate pharmacovigilance.	Drug safety	BilSTM[kk], CRF-NN[ll]	Joint modeling improved the identification of adverse drug events from 0.62 to 0.65
Dey et al [103]	To predict and prevent adverse drug reactions at an early stage to enhance drug safety	Drug safety	Deep learning	Neural fingerprints from the deep learning model (AUC=0.72) outperformed all other methods in predicting adverse drug reactions. The model identified important molecular substructures that are associated with specific adverse drug reactions
Yang et al [104]	To identify medications, adverse drug effects, and their relations with clinical notes	Drug safety	MADEx, LSTM-RNN[mm], CRF[nn], SVM, RF	MADEx achieved the top-three best performances (*F1* score of 0.8233) for clinical name entity recognition, adverse drug effect, and relations from clinical texts, which outperformed traditional methods
Chapman et al [105]	To identify adverse drug effect symptoms and drugs in clinical notes	Drug safety	NLP	The micro-averaged *F1* score was 80.9% for named entity recognition, 88.1% for relation extraction, and 61.2% for the integrated systems
Lian et al [106]	To detect adverse drug reactions	Drug safety	LRM[oo], BNM[pp], BCP-NN[qq]	Experimental results showed the usefulness of the proposed pattern discovery method by improving the standard baseline adverse drug reaction by 23.83%

Reference	Objective	Study theme	AI[a] method	Findings (patient safety outcomes)
Huang et al [107]	To predict adverse drug effects	Drug safety	SVM, LR	The proposed computational framework showed that an in silico model built on this framework can achieve satisfactory cardiotoxicity adverse drug reaction prediction performance (median AUC=0.771, accuracy=0.675, sensitivity=0.632, and specificity=0.789).

[a]AI: artificial intelligence.
[b]KNN: K-nearest neighbor.
[c]NB: naive Bayes.
[d]LR: logistic regression.
[e]SVM: support vector machine.
[f]RF: random forest.
[g]AUC: area under the curve.
[h]ICU: intensive care unit.
[i]MMD: multimodal section.
[j]DT: decision tree.
[k]BCT: binary classification tree.
[l]RDAC: regularized discriminant analysis classifier.
[m]NN: neural network.
[n]DEWS: deep learning–based early warning system.
[o]MEWS: modified early warning system.
[p]AUPRC: area under the precision-recall curve.
[q]J48: decision tree algorithm.
[r]CC: closure classifier.
[s]MLP: multilayer perceptron.
[t]NLP: natural language processing.
[u]CNN: convolutional neural network.
[v]ET: extra tree.
[w]RNN: recurrent neural network.
[x]NCPS: National Center for Patient Safety.
[y]XGB: extreme gradient boosting.
[z]LASSO: least absolute shrinkage and selection operator.
[aa]HIT: health information technology.
[bb]PSE: patient safety event.
[cc]PANDIT: Patient Assisting Net-Based Diabetes Insulin Titration.
[dd]CHAID: Chi square automatic interaction detector.
[ee]CART: classification and regression tree.
[ff]SVR: support vector regression.
[gg]NN-BP: neural network-back propagation.
[hh]MT: model tree.
[ii]ABC4D: Advanced Bolus Calculator For Diabetes.
[jj]CSS: clinical support system.
[kk]BiLSTM: bi-long short-term memory neural network.
[ll]CRF-NN: conditional random field neural network.
[mm]LSTM-RNN: long short-term memory-recurrent neural network.
[nn]CRF: conditional random field neural network.
[oo]LRM: logistic regression probability model.
[pp]BNM: Bayesian network model.
[qq]BCP-NN: Bayesian confidence propagation neural network.

Table 2. Performance of artificial intelligence.

Reference	Best model recommended	Comparison/other models	Performance measures of the best model						
			Accuracy	AUROC[a]	Recall	Specificity	Precision	*F* measure	other
Huanget al [107]	SVM[b]	Logistic regression	0.675	0.771	0.632	0.789	N/A[c]	N/A	N/A
Lian et al [106]	Ensemble of three models	Bayesian network model; likelihood ratio model; BCPNN[d]	N/A	N/A	N/A	N/A	N/A	N/A	Chi-square improved by 28.83%
Chapman et al [105]	Integrated NLP[e] with RF[f] model for relation extraction and CRF[g] model	CRF; RF model for relation extraction	N/A	N/A	N/A	N/A	N/A	0.612	N/A
Yang et al [104]	MADEx (long short-term memory CRF+SVM)	RNN[h]; CRF; SVM; RF	N/A	N/A	0.6542	N/A	0.5758	0.6125	N/A
Dey et al [103]	Neural fingerprint (deep learning)	10 other chemical fingerprints	0.91	0.82	0.50	0.93	N/A	0.400	N/A
Dandala et al [102]	BiLSTM[i]+CRF (joint and external resources)	BiLSTM+CRF (sequential); BiLSTM+CRF (joint)	N/A	N/A	0.822 concept extraction; 0.855 relation classification	N/A	0.846 concept extraction; 0.888 relation classification	0.83 concept extraction; 0.87 relation classification	N/A
Cai et al [101]	CARD[j]	Association rule mining	N/A	N/A	N/A	N/A	N/A	N/A	Identifying drug interaction 20%
Onay et al [100]	LSVM[k]	Boosted and bagged trees (ensemble)	0.89	0.88	0.83	1.00	N/A	0.91	N/A
Tinoco et al [99]	Computerized surveillance system	Manual chart review	N/A	N/A	N/A	N/A	N/A	N/A	Number of events detected 92% (HAI[l]), 82% (SSI[m]), 91% (LR-TI[n]), 99% (UTI[o]), 100% (BSI[p]), 52% (ADE[q])

Reference	Best model recommended	Comparison/other models	Performance measures of the best model						
			Accuracy	AUROC[a]	Recall	Specificity	Precision	*F* measure	other
Carrel et al [98]	NLP-assisted manual review	Manual chart review	N/A	N/A	N/A	N/A	N/A	N/A	Identified 3.1% additional patients with opioid problems
Li et al [97]	NLP-based hybrid model	Rule-based method; CRF	N/A	N/A	0.907	N/A	0.924	0.915	N/A
Schiff et al [96]	MedAware, a probabilistic machine-learning CDS[r] system	Traditional CDS	0.75	N/A	N/A	N/A	N/A	N/A	75% of the identified alerts were clinically meaningful
Reddy et al [95]	ABC4D[s] smartphone app (based on CBR[t], an AI[u] technique)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	ABC4D was superior to nonadaptive bolus calculator and also more user friendly
Long et al [93]	AI smartphone app	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100% adherence in the intervention group
Hasan et al [92]	Co-occurrence KNN[v] and popular algorithm	Logistic regression; KNN; random algorithm; co-occurrence; drug popularity	N/A	N/A	N/A	N/A	N/A	N/A	Simple algorithms such as popular algorithm, co-occurrence, and KNN performed better than more complex logistic regression

Reference	Best model recommended	Comparison/other models	Performance measures of the best model						
			Accuracy	AUROC[a]	Recall	Specificity	Precision	*F* measure	other
Hu et al [91]	Bagged SVR[w] and bagged voting	MLP[x]; model tree; KNN	N/A	N/A	N/A	N/A	N/A	N/A	Mean absolute error for both 0.210
Tang et al [90]	NLP	N/A	N/A	N/A	0.59	N/A	0.75	N/A	N/A
Hu et al [89]	RF	C4.5; KNN; CART[y]; MLP; logistic regression	0.839	0.912	0.782	0.888	N/A	N/A	N/A
Bean et al [88]	Own model	Logistic regression; SVM; decision tree; NLP	N/A	0.92	N/A	N/A	N/A	N/A	N/A
Hamma et al [87]	CART	CART and CHAID[z]	0.902	N/A	N/A	N/A	N/A	N/A	CHAID outperformed CART only in central nervous system classification
Song et al [86]	Similarity-based SVM	Analogous machine-learning algorithms (not mentioned)	N/A	N/A	0.24	0.97	0.68	N/A	N/A
Simon et al [85]	PANDIT[aa]	Nurses	0.635	N/A	N/A	N/A	N/A	N/A	36.5% PANDIT recommendation did not match with the nurses; 1.4% of the recommendations were unsafe.
Fong et al [84]	Unigram logistic regression	Unigram, bigram, and combined logistic regression and SVM	N/A	0.914	0.830	N/A	0.838	0.765	Unigram SVM and logistic regression were comparable
Ye et al [83]	RF	Linear and nonlinear machine-learning algorithms	N/A	N/A	N/A	N/A	N/A	N/A	C-statistic of 0.884
Marella et al [82]	Naïve Bayes kernel	Naïve Bayes; KNN and rule induction	0.855	0.927	N/A	N/A	N/A	0.877	N/A
McKnight [81]	NLP; SELF[bb]	N/A	Labeled 0.52; unlabeled 0.80	N/A	N/A	N/A	N/A	N/A	N/A

Reference	Best model recommended	Comparison/other models	Performance measures of the best model						
			Accuracy	AUROC[a]	Recall	Specificity	Precision	*F* measure	other
Rosenbaum and Baron [80]	SVM	Logistic regression	N/A	0.97	0.80	0.96	N/A	N/A	Positive predictive value 0.52
Wang et al [79]	Binary SVM with radial basis function kernel	Regularized logistic regression; linear SVM	N/A	N/A	0.783	N/A	0.783	0.783	N/A
Gupta and Patrick [66]	Naïve Bayes multinomial	J48; naïve Bayes; SVM	N/A	0.96	0.78	0.98	0.79	0.78	Kappa 0.76; mean absolute error 0.03
Wang et al [67]	Ensemble classifier chain of SVM with radial basis function kernel	Binary relevance of SVM, classifier chain of SVM	0.654	N/A	0.791	N/A	0.689	0.736	Hamming loss 0.80
Zhou et al [49]	SVM and RF	Naïve Bayes and MLP	N/A	N/A	0.769 SVM for event type; 0.927 RF for even cause	N/A	0.788 SVM for event type; 0.927 RF for event cause	0.758 SVM for event type; 0.925 RF for event cause	N/A
Fong et al [68]	NLP with SVM	NLP with decision tree	0.990	0.960	0.920	1.00	1.000	0.960	N/A
El Messiry et al [69]	NLP	Scaled linear discriminant analysis; SVM; LASSO[cc] and elastic-net regularized generalized linear models; max entropy; RF; neural network	0.730	N/A	0.770	0.696	N/A	N/A	N/A
Chondrogiannis et al [70]	NLP	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Model developed in this study identified that each clinical report contains about 6.8 abbreviations
Liang and Gong [71]	Naïve Bayes with binary relevance	SVM; decision rule; decision tree; KNN	N/A	N/A	N/A	N/A	N/A	N/A	Micro F measure 0.212

Reference	Best model recommended	Comparison/other models	Performance measures of the best model						
			Accuracy	AUROC[a]	Recall	Specificity	Precision	*F* measure	other
Ong et al [72]	Text classifier with SVM	Text classifier with naïve Bayes	N/A	0.920 multitype dataset; 0.980 patient misidentification dataset	0.830 multitype dataset; 0.940 patient misidentification dataset	N/A	0.880 multitype dataset; 0.990 patient misidentification dataset	0.860 multitype dataset; 0.960 patient misidentification dataset	N/A
Taggart et al [73]	Rule-based NLP	SVM; extra trees; convolutional neural network	N/A	N/A	N/A	0.846	N/A	N/A	Positive predictive value 0.627; negative predictive value 0.971
Denecke et al [74]	AIML[dd]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Minimize information loss during clinical visits
Evans et al [75]	SVM	J48; naïve Bayes	0.728	0.891 incident type; 0.708 severity of harm	N/A	N/A	N/A	N/A	N/A
Wang et al [76]	Convolutional neural network	SVM	N/A	N/A	N/A	N/A	N/A	0.850	N/A
Klock et al [47]	SVM and RNN[ee]	RF	0.899 SVM; 0.900 RNN	N/A	N/A	N/A	N/A	0.648899 SVM; 0.889 RNN	N/A
Li et al [77]	Ensemble machine learning (bagging, boosting, and random feature method)	N/A	N/A	N/A	0.572 from 0.10 risk score; 0.855 from 0.04 risk score	N/A	N/A	N/A	C-statistic 0.880
Muff et al [78]	NLP	Patient safety indicators	N/A	N/A	0.770	0.938	N/A	N/A	N/A
Kwon et al [65]	Deep learning-based early warning system	Modified early warning system; RF; logistic regression	N/A	0.850	0.757	0.765	N/A	1.000	AUPRC[ff]
Hu et al [64]	Neural network model	ViEWS[gg]	N/A	0.880	N/A	N/A	N/A	0.81	Positive predictive value 0.726

Reference	Best model recommended	Comparison/other models	Performance measures of the best model						
			Accuracy	AUROC[a]	Recall	Specificity	Precision	*F* measure	other
Segal et al [63]	MedAware (a CDSS[hh]) + EHR[ii]	Legacy CDS	N/A	N/A	N/A	N/A	N/A	N/A	Clinically relevant 85%, alert burden 0.04%
Menard et al [62]	Machine learning (name not disclosed)	N/A	N/A	0.970	N/A	N/A	N/A	N/A	N/A
Eerikainen et al [61]	RF	Binary classification tree; regularized discriminant analysis classifier; SVM; RF	N/A	N/A	0.950	0.780	N/A	0.782	N/A
Antink et al [60]	Combined (selecting the best machine- learning algorithm for each alarm type)	Binary classification tree; regularized discriminant analysis classifier; SVM; RF	N/A	N/A	0.950	0.780	N/A	0.782	N/A
Zhang et al [59]	Cost-sensitive SVM	N/A	N/A	N/A	0.950	0.850	N/A	0.809	N/A
Ansari et al [58]	Multimodal machine learning using decision tree	N/A	N/A	N/A	0.890	0.850	N/A	0.762	N/A

Reference	Best model recommended	Comparison/other models	Performance measures of the best model						
			Accuracy	AUROC[a]	Recall	Specificity	Precision	*F* measure	other
Chen et al [57]	RF	N/A	N/A	0.870	N/A	N/A	N/A	N/A	N/A

[a]AUROC: area under the receiver operating characteristic curve.
[b]SVM: support vector machine.
[c]N/A: not applicable (Not reported).
[d]BCPNN: Bayesian confidence propagation neural network.
[e]NLP: natural language processing.
[f]RF: random forest.
[g]CRF: conditional random field.
[h]RNN: recurrent neural network.
[i]BiLSTM: Bi-long short-term memory neural network.
[j]CARD: casual association rule discovery.
[k]LSVM: linear support vector machine.
[l]HAI: hospital-associated infection.
[m]SSI: surgical site infection.
[n]LRTI: lower respiratory tract infection.
[o]UTI: urinary tract infection.
[p]BSI: bloodstream infection.
[q]ADE: adverse drug event.
[r]CDS: clinical decision support.
[s]ABC4D: Advanced Bolus Calculator For Diabetes.
[t]CBR: case-based reasoning.
[u]AI: artificial intelligence.
[v]KNN: K-nearest neighbor.
[w]SVR: support vector regression.
[x]MLP: multilayer perceptron.
[y]CART: classification and regression tree.
[z]CHAID: Chi square automatic interaction detector.
[aa]PANDIT: Patient Assisting Net-Based Diabetes Insulin Titration.
[bb]SELF: semisupervised local Fisher discriminant analysis.
[cc]LASSO: least absolute shrinkage and selection operator.
[dd]AIML: artificial intelligence markup language.
[ee]RNN: recurrent neural network.
[ff]AUPRC: area under the precision-recall curve.
[gg]ViewS: VitalPac Early Warning Score.
[hh]CDSS: clinical decision support system.
[ii]EHR: electronic health record.

Study Themes and Findings

Clinical Alarms and Alerts

Nine publications addressed clinical alarms/alerts using AI techniques. The most widely used method was random forest (n=5) followed by support vector machine (n=3) and neural network/deep learning (n=3).

Studies under this category used electrocardiogram data from the PhysioNet Challenge public database and PhysioNet MIMIC II database. Five studies focused on reducing false alarm rates arising due to cardiac ailments such as arrhythmia and cardiac arrest in an intensive care unit setting [58-61,65]. The remaining four studies focused on improving the performance of clinical alarms in classifying clinical deterioration such as fluctuation in vital signs [57], predicting adverse events [62], identifying adverse medication events [63], and deterioration of patient health with hematologic malignancies [64].

Clinical Reports

We identified 21studies concerning clinical reports. Studies in this group primarily focused on extracting information from clinical reports such as safety reports (internal to the hospital), patient feedback, EHR notes, and others typically derived from incident monitoring systems and patient safety organizations. The most widely used method was support vector machine (n=11), followed by natural language processing (n=7) and naïve Bayes (n=5). We also identified decision trees (n=4), deep learning models (n=3), J48 (n=2), and other (n=9) algorithms.

The majority of articles focused on automating the process of patient safety classifications. These studies used machine learning and natural language processing techniques to classify clinical incidents [66] from the Incident Information Management System and to identify risky incidents [71,79,81,108] in patient safety reports retrieved from different sources, including the university database and the Veterans Affairs National Center for Patient Safety database. Some studies also analyzed medication reports [49] from structured and unstructured data obtained from the patient safety organization, and evaluated patient feedback [69] retrieved from the Patient Advocacy Reporting System developed at Vanderbilt and associated institutions.

Several studies focused on classifying the type and severity of patient safety incident reports using data collected by different sources such as universities [75], and incident reporting systems such as Advanced Incident Management Systems (across Australia) and Riskman [67,75,76]. Others analyzed hospital clinical notes internally (manually annotated by clinicians and a quality committee) and data retrieved from patient safety organizations to identify adverse incidents such as delayed medication [68], fall risks [47,67], near misses, patient misidentification, spelling errors, and ambiguity in clinical notes [109]. One study analyzed clinical study descriptions from clinicaltrials.gov and implemented an AI system to detect all abbreviations and identify their meaning to minimize incorrect interpretations [70]. Another study used inpatient laboratory test reports from Sunquest Laboratory Information System and identified wrong blood in tube errors [80].

Studies used clinical reports from various sources, including patient safety organizations, EHR data from Veterans Health Administration and Berkshire Health Systems, and deidentified notes from the Medical Information Mart for Intensive Care. These studies focused on extracting relevant information [74,77,82,84] to predict bleeding risks among critically ill patients [73], postoperative surgical complications [78], mortality risk [83], and other factors such as lab test results and vital signs [77] influencing patient safety outcomes.

Adverse Drug Events or Drug Safety

Twenty-three publications were classified under drug safety. These studies primarily addressed adverse effects related to drug reactions. The most widely used method was random forest (n=8), followed by natural language processing (n=7) and logistic regression (n=6). Algorithms including natural language processing (n=5), logistic regression (n=4), mobile or web apps (n=3), AI devices (n=2), and others (n=5) were also used.

Studies in this category retrieved data from different repositories such as DrugBank, Side Effect Resource, the Food and Drug Administration (FDA)'s adverse event reporting system, University of Massachusetts Medical School, Observational Medical Outcomes Partnership database, and Human Protein-Protein Interaction database to identify adverse drug interactions and reactions that can potentially negatively influence patient health [86-88,101,102,105-107,110]. Some studies also used AI to predict drug interactions by analyzing EHR data [88], unstructured discharge notes [90], and clinical charts [99,104]. One study also used AI to identify drugs that were withdrawn from the commercial markets by the FDA [100].

Some studies used AI to predict the dosage of medicines such as insulin, digoxin, and warfarin [85,89,91,95]. AI in drug safety was also used to scan through the hospital's EHR data and identify medication errors (ie, wrong medication prescriptions) [96]. One study used AI to monitor stroke patients and track their medication (anticoagulation) intake [93]. Several studies used AI to predict a medication that a patient could be consuming but was missing from their medication list or health records [92,94,97]. Another study used AI to review clinical notes and identify evidence of opioid abuse [98].

Visual Representations of Safety and Chronology of the Studies

Figure 5 illustrates the details of patient safety issues/outcomes studied and reported under each classified theme using AI algorithms at the clinical level.

Figure 5. Identified factors influencing patient safety outcomes. EHR: electronic health record.

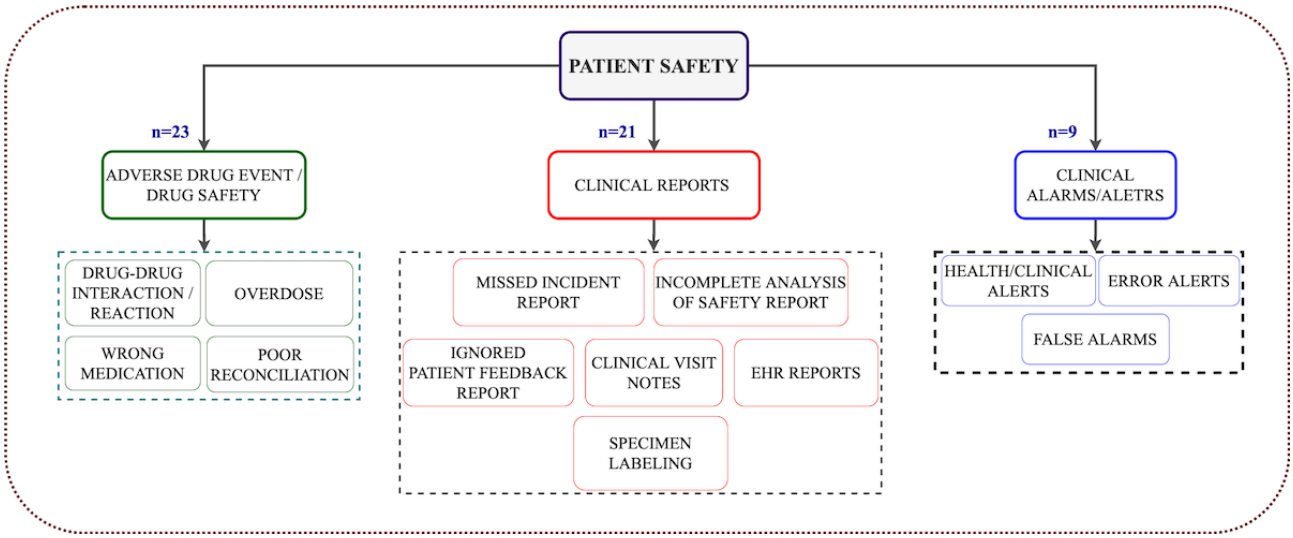
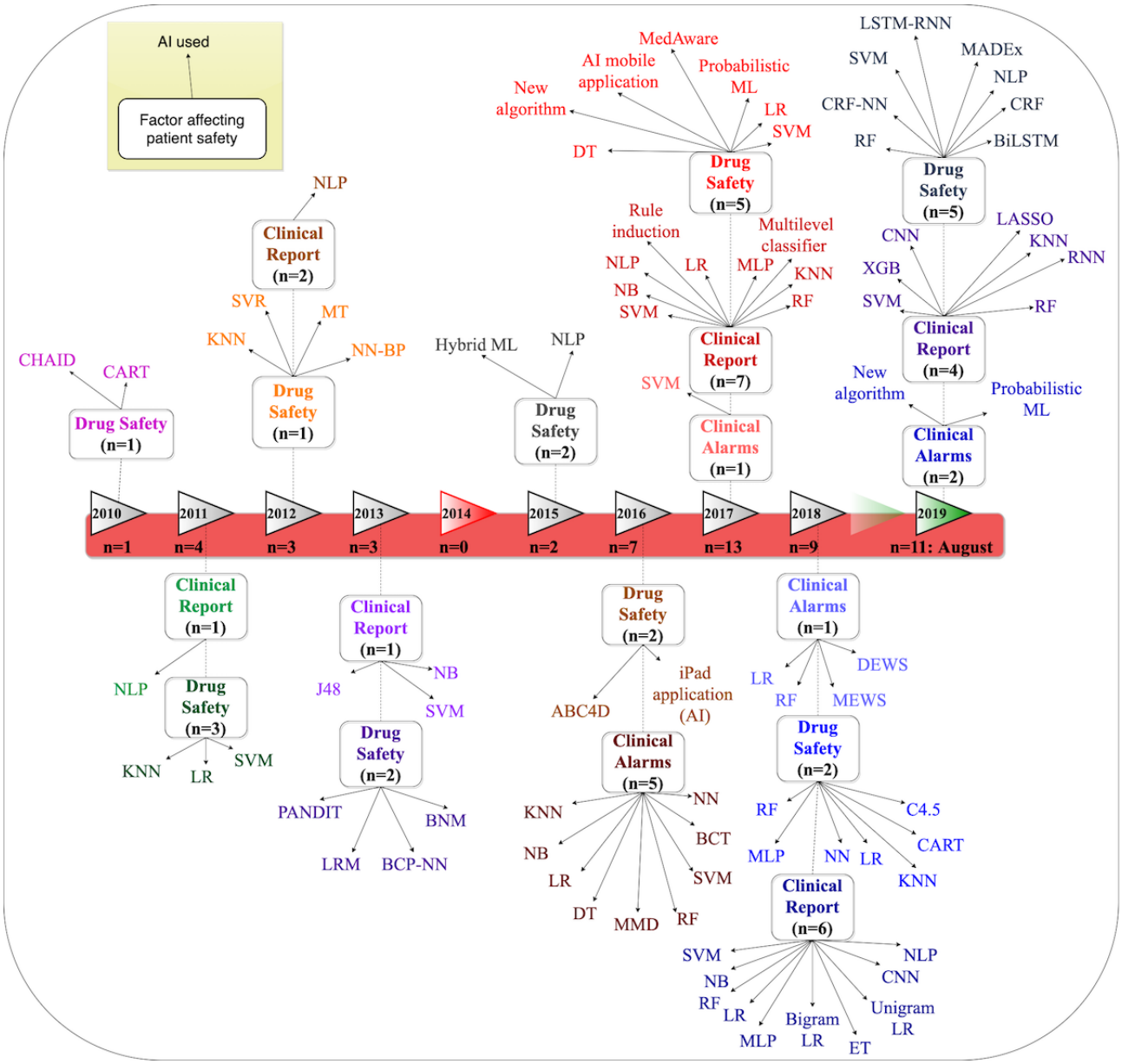


Figure 6 further shows how the application of AI in studies reporting patient safety outcomes in our review evolved over time between January 2009 and August 2019.

Figure 6. Timeline of artificial intelligence application to address factors influencing patient safety (clinical reports, drug safety, and clinical alarms) between 2009 and August 2019. ABC4D: Advanced Bolus Calculator For Diabetes; AI: artificial intelligence; BCP-NN: Bayesian confidence propagation neural network; BCT: binary classification tree; BiLSTM: bi-long short-term memory neural network; BNM: Bayesian network model; CART: classification and regression tree; CHAID: Chi-square automatic interaction detector; CRF-NN: conditional random field neural network; DEWS: deep learning-based early warning system; DT: decision tree; KNN, K-nearest neighbor; LASSO: least absolute shrinkage and selection operator; LR: logistic regression; LSTM-RNN: long short-term memory-recurrent neural network; MEWS: modified early warning system; ML: machine learning; MLP: multilayer perception; MMD; multimodal detection; MT: model tree; NB: naive Bayes; NLP: natural language processing; NN: neural network; NN-BP: neural network back propagation; PANDIT: Patient Assisting Net-Based Diabetes Insulin Titration; RF: random forest; RNN: recurrent neural network; SVM: support vector machine; SVR, support vector regression; XGB; extreme gradient boosting.



Discussion

Principal Findings

Many studies have been conducted to exhibit the analytical performance of AI in health care, particularly as a diagnostic and prognostic tool. To our knowledge, this is the first systematic review exploring and portraying studies that show the influence of AI (machine-learning and natural language processing techniques) on clinical-level patient safety outcomes.

We identified 53 studies within the scope of the review. These 53 studies used 38 different types of AI systems/models to address patient safety outcomes, among which support vector machine (n=17) and natural language processing (n=12) were the most frequently used. Most of the reviewed studies reported positive changes in patient safety outcomes.

Analysis of all studies showed that there is a lack of a standardized benchmark among reported AI models. Despite varying AI performance, most studies have reported a positive impact on safety outcomes (Table 2), thus indicating that safety

outcomes do not necessarily correlate to AI performance measures [26]. For example, one identified study with an accuracy of 0.63 that implemented Patient Assisting Net-Based Diabetes Insulin Titration (PANDIT) reported a negative impact of AI on safety outcomes. The PANDIT-generated recommendations that did not match with the recommendations of nurses (1.4% of the recommendations) were identified as unsafe [85]. In contrast, the study implementing natural language processing to extract clinical information from patient safety reports showed a positive impact on patient safety outcomes with accuracy of 0.53 [81]. Similarly, the FDA-approved computer-aided diagnosis of the 1990s, which significantly increased the recall rate of diagnosis, did not improve safety or patient outcomes [111]. According to our review, AI algorithms are rarely scrutinized against a standard of care (clinicians or clinical gold standard). Relying on AI outcomes that have not been evaluated against a standard benchmark that meets clinical requirements can be misleading. A study conducted in 2008 [112] developed and validated an advanced version of the QRISK cardiovascular disease risk algorithm (QRISK2). The study reported improved performance of QRISK2 when compared to its earlier version. However, QRISK2 was not compared against any clinical gold standard. Eight years later, in 2016, The Medicines & Healthcare Products Regulatory Agency identified an error in the QRISK 2 calculator [113]; QRISK2 underestimated or overestimated the potential risk of cardiovascular disease. The regulatory agency reported that a third of general practitioner surgeries in England might have been affected [113] due to the error in QRISK2. Globally, there are several Standards Development Organizations developing information technology and AI standards to address varying standardization needs in the domain of cloud computing, cybersecurity, and the internet of things [114]. However, there has been minimal effort to standardize AI in the field of health care. Health care comprises multiple departments, each having unique or different requirements (clinical standards). Thus, health care requires so-called “vertical standards,” which are standards developed for specific application areas such as drug safety (pharmaceuticals), specific surgeries, outpatients and inpatients with specific health concerns, and emergency departments [114]. In contrast, standards that are not correctly tailored for a specific purpose may hamper patient safety.

Without a standardized benchmark, it becomes challenging to evaluate whether a particular AI system meets clinical requirements (gold standard) or performs significantly better (improves patient safety) or worse (harms patient) than other similar systems in a given health care context. To generate the best possible (highest) performance outcome, AI algorithms may include unreliable confounders into the computing process. For instance, in one study, an algorithm was more likely to classify a skin lesion as malignant if an image (input data) had a ruler in it because the presence of a ruler correlated with an increased likelihood of a cancerous lesion [115]. The presence of surgical skin markings has also been shown to falsely increase a deep-learning model’s melanoma probability scores and hence the false-positive rate [116]. Moreover, there has been great emphasis focused on the importance to standardization of AI by developed countries such as the European Union, United States, China, and Japan. For instance, on February 11, 2019, the President of the United States issued an Executive Order (EO 13859) [117] directing federal agencies to actively participate in AI standards development. According to the Center for Data Innovation and The National Institute of Standards and Technology, a standardized AI benchmark can serve as a mechanism to evaluate and compare AI systems [114]. FDA Commissioner Scott Gottlieb acknowledged the importance of AI standardization that can assure that ongoing algorithm changes follow prespecified performance objectives and use a validation process that ensures safety [118].

Another major finding of this review is high heterogeneity in AI reporting. AI systems have been developed to help clinicians in estimating risks and making informed decisions. However, the evidence indicates that the quality of reporting of AI model studies is heterogeneous (not standard). Table 2 demonstrates how different studies that implemented the same AI used different evaluation metrics to measure its performance. Heterogeneity in AI reporting also makes the comparison of algorithms across studies challenging and might cause difficulties in obtaining consensus while attempting to select the best AI for a given situation. Algorithms not only need to be subjected to comparison on the same data that are representative of the target population but also the same evaluation metrics; thus, standardized reporting of AI studies would be beneficial. The current Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) consists of 22-item checklists that aim to improve the reporting of studies developing or validating a prediction model [119,120]. Studies in our review did not use TRIPOD to report findings. The possible reason behind this can be the design of TRIPOD, which focuses on a regression-based prediction model.

However, the explanation and elaboration document provides examples of good reporting methods, which are focused on models developed using regression. Therefore, a new version of the TRIPOD statement that is specific to AI/machine-learning systems (TRIPOD-ML) is in development. It will focus on the introduction of machine-learning prediction algorithms to establish methodological and reporting standards for machine-learning studies in health care [121].

Our findings also identified the need to determine the importance of an AI evaluation metric. In particular, it is important to determine which evaluation metric(s) should be measured in a given health care context. AUROC is considered to be a superior metric for classification accuracy, particularly when unbalanced datasets are used [122,123] because it is unaffected by unbalanced data, which is typical in health care. However, 36 studies in our review did not report AUROC. Evaluation measures such as precision-recall can also reflect model performance accurately [123]; however, only 11 studies in our review evaluated AI based on precision-recall. Using inappropriate measures to evaluate AI performance might impose a threat to patient safety. However, no threat to patient safety due to the use of inappropriate AI evaluation metric was identified in our review. Future studies should report the importance of evaluation metrics and determine which measure (single or multiple measures) is more important and a better representation of patient safety outcomes. More studies are

needed to explore the evaluation metric(s) that should be considered before recommending an AI model.

The findings of our review demonstrate that drug safety, followed by the analysis of clinical reports, has been the most common area of interest for the use of AI to address clinical-level patient safety concerns. The wrong medication or improper dosage can result in fatal patient health outcomes and medical malpractice [91]. Of all drug safety concerns, issues related to inappropriate doses of high-alert medications are of great interest to the Joint Commission on Accreditation of Healthcare Organizations [91,124]. Medical errors are reported as the third leading cause of death in the United States. The majority of the papers in our review implemented AI to address drug safety (n=23) concerns, which is one of the most significant contributors to overall medical errors. These publications improved patient safety by identifying adverse drug reactions and preventing incorrect medications or overdoses. Future studies should further explore how to use AI systems on a larger scale to diminish medication errors at hospitals and clinics to save more lives.

Finally, the studies reviewed in this paper have addressed safety issues as identified by the Health Insurance Portability and Accountability Act (HIPAA) and the US Department of Health & Human Services (HHS). The HIPAA regulations identify risk analysis as part of the administrative safeguard requirement to improve patient safety. The HHS advocates analysis of clinical notes to track, detect, and evaluate potential risks to patients. Many studies (n=21) in our review used AI to identify patient risk from clinical notes. These studies used AI and clinical reports to extract safety-related information such as fall risks, pyxis discrepancies, patient misidentification, patient severity, and postoperative surgical complications. Our findings exhibit how, with the help of AI techniques such as natural language processing, clinical notes and reports have been used as a data source to extract patient data regarding a broad range of safety issues, including clinical notes, discharge notes, and other issues [69,70,73,84]. Our review also indicates that AI has the potential to provide valuable insights to treat patients correctly by identifying future health or safety risks [125], to improve health care quality, and reduce clinical errors [126]. Despite being recognized as one of the major factors responsible for fatigue, burnout in clinicians, and patient harm [61,127-129], only 9 studies in our review used AI to improve clinical alarms. Although studies addressing clinical alarms reported positive outcomes by minimizing false alarms and identifying patient health deterioration, the limited number of studies (n= 9) addressing these issues shows that the field is still in a nascent period of investigation. Thus, more research is needed to confirm the impact of AI on patient safety outcomes.

Recommendations for Future Research

Future studies should work toward establishing a gold standard (for various health care contexts/ disease types/problem types) against which AI performance can be measured. Future research, as suggested by Kelly and others in 2019 [119], should also develop a common independent test (preferably for different problem types, drug safety/clinical alarms/clinical reports) using unenriched representative sample data that are not available to train algorithms.

Our review acknowledges that no single measure captures all of the desirable properties of a model, and multiple measures are typically required to summarize model performance. However, different measures are indicative of different types of analytical performance. Future studies should develop a standard framework that can guide clinicians in interpreting the clinical meaning of AI's evaluation metrics before integrating it into the clinical workflow. Future studies should also report a quantifiable measure of AI demonstrating not only its analytical performance but also its impact on patient safety (long and short term), reliability, domain-specific risks, and uncertainty. Additionally, studies should also ensure data standardization.

Health databases or storage systems are often not compatible (integratable) across different hospitals, care providers, or different departments in the same hospital. Data in health care are largely unorganized and unstructured [9,50]. Since the performance of AI heavily depends on data, regulatory bodies should invest in data infrastructure such as standardization of EHRs and integration of different health databases. AI trained on unstructured or biased data might generate misleading results [51]. According to the National Institute of Standards and Technology (NIST), standardized data can make the training data (machine learning input) more visible and usable to authorized users. It can also ensure data quality and improve AI performance.

Most of the safety initiatives implemented in health care over the last decade have been focused on analyzing historical events to learn and evolve [130,131]. The same was also observed in our review. AI models were trained on past data. However, in health care, outcomes are satisfactory because providers make sensible and just-in-time adjustments according to the demands of the situation. Future work should train AI on the critical adjustments made by clinicians, so that AI can adapt to different conditions in the same manner as clinicians.

The integration of AI systems into the health system will alter the role of providers. Ideally, AI systems are expected to assist providers in making faster and more accurate decisions and to deliver personalized patient care. However, lack of appropriate knowledge of using complex AI systems and interpreting their outcome might impose a high cognitive workload on providers. Thus, the medical education system should incorporate necessary AI training for providers so that they can better understand the basic functioning of AI systems and extract clinically meaningful insight from the outcomes of AI.

Limitation of this Review

This study encompasses publications that matched our inclusion criteria and operational definition of AI and patient safety. In addition, we limited the scope of AI to only machine learning and natural language processing at a clinical level. This review also only included studies published in English in the last 10 years.

Conclusion

This systematic review identified critical research gaps that need attention from the scientific community. The majority of the studies in the review have not highlighted significant aspects of AI, such as (a) heterogeneity in AI reporting, (b) lack of a standardized benchmark, and (c) need to determine the importance of AI evaluation metric. The identified flaws of AI systems indicate that further research is needed, as well as the involvement of the FDA and NIST to develop a framework standardizing AI evaluation measures and set a benchmark to ensure patient safety. Thus, our review encourages the health care domain and AI developers to adopt an interdisciplinary and systems approach to study the overall impact of AI on patient safety outcomes and other contexts in health care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) checklist.
[DOC File , 65 KB - medinform_v8i7e18599_app1.doc]

References

1. World Health Organization. Regional strategy for patient safety in the WHO South-East Asia Region (2016-2025). India: WHO Regional Office for South-East Asia; 2015.
2. Minimal Information Model for Patient Safety Incident Reporting and Learning Systems. World Health Organization. 2016. URL: https://www.who.int/patientsafety/topics/reporting-learning/mim/user-guide/en/ [accessed 2020-06-20]
3. World Health Organization. Provisional agenda item 12.5. Provisional agenda item 12. 2019 Presented at: Seventy-seccond World Health Assembly; May 20-28, 2019; Palais des Nations, Geneva URL: https://apps.who.int/gb/ebwha/pdf_files/WHA72/A72_JOUR3-en.pdf
4. Liang C, Miao Q, Kang H, Vogelsmeier A, Hilmas T, Wang J, et al. Leveraging Patient Safety Research: Efforts Made Fifteen Years Since To Err Is Human. Stud Health Technol Inform 2019 Aug 21;264:983-987. [doi: 10.3233/SHTI190371] [Medline: 31438071]
5. James JT. A new, evidence-based estimate of patient harms associated with hospital care. J Patient Saf 2013 Sep;9(3):122-128. [doi: 10.1097/PTS.0b013e3182948a69] [Medline: 23860193]
6. Macrae C. Governing the safety of artificial intelligence in healthcare. BMJ Qual Saf 2019 Jun;28(6):495-498. [doi: 10.1136/bmjqs-2019-009484] [Medline: 30979783]
7. Grossman L, Choi S, Collins S, Dykes P, O'Leary K, Rizer M, et al. Implementation of acute care patient portals: recommendations on utility and use from six early adopters. J Am Med Inform Assoc 2018 Apr 01;25(4):370-379. [doi: 10.1093/jamia/ocx074] [Medline: 29040634]
8. McCarthy J, Hayes P. Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer B, Michie D, editors. Machine Intelligence 4. Edinburgh: Edinburgh University Press; 1969:463-502.
9. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. Ann Surg 2018 Jul;268(1):70-76 [FREE Full text] [doi: 10.1097/SLA.0000000000002693] [Medline: 29389679]
10. Bhardwaj R, Nambiar A, Dutta D. A Study of Machine Learning in Healthcare. 2017 Presented at: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC); July 4-8, 2017; Turin, Italy. [doi: 10.1109/COMPSAC.2017.164]
11. Kong H. Managing Unstructured Big Data in Healthcare System. Healthc Inform Res 2019 Jan;25(1):1-2 [FREE Full text] [doi: 10.4258/hir.2019.25.1.1] [Medline: 30788175]
12. Lee R, Lober W, Sibley J, Kross E, Engelberg R, Curtis J. Identifying Goals-of-Care Conversations in the Electronic Health Record Using Machine Learning and Natural Language Processing. Am J Resp Crit Care 2020;201(1):A1089. [doi: 10.1164/ajrccm-conference.2019.199.1_meetingabstracts.a1089]
13. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
14. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
15. Lu H, Li Y, Chen M, Kim H, Serikawa S. Brain Intelligence: Go beyond Artificial Intelligence. Mobile Netw Appl 2017 Sep 21;23(2):368-375. [doi: 10.1007/s11036-017-0932-8]
16. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision. Radiology 2018 Mar;286(3):810-818. [doi: 10.1148/radiol.2017170549] [Medline: 29039725]
17. Guan M, Cho S, Petro R, Zhang W, Pasche B, Topaloglu U. Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes. JAMIA Open 2019 Apr;2(1):139-149 [FREE Full text] [doi: 10.1093/jamiaopen/ooy061] [Medline: 30944913]

18. Li Q, Zhao K, Bustamante CD, Ma X, Wong WH. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. Genet Med 2019 Sep;21(9):2126-2134 [FREE Full text] [doi: 10.1038/s41436-019-0439-8] [Medline: 30675030]
19. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018 Jun;23(6):1241-1250 [FREE Full text] [doi: 10.1016/j.drudis.2018.01.039] [Medline: 29366762]
20. Sahli Costabal F, Matsuno K, Yao J, Perdikaris P, Kuhl E. Machine learning in drug development: Characterizing the effect of 30 drugs on the QT interval using Gaussian process regression, sensitivity analysis, and uncertainty quantification. Comput Meth Appl Mechan Eng 2019 May;348:313-333. [doi: 10.1016/j.cma.2019.01.033]
21. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater 2019 May;18(5):435-441 [FREE Full text] [doi: 10.1038/s41563-019-0338-z] [Medline: 31000803]
22. Banerjee I, Li K, Seneviratne M, Ferrari M, Seto T, Brooks JD, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMIA Open 2019 Apr;2(1):150-159 [FREE Full text] [doi: 10.1093/jamiaopen/ooy057] [Medline: 31032481]
23. Ciervo J, Shen SC, Stallcup K, Thomas A, Farnum MA, Lobanov VS, et al. A new risk and issue management system to improve productivity, quality, and compliance in clinical trials. JAMIA Open 2019 Jul;2(2):216-221 [FREE Full text] [doi: 10.1093/jamiaopen/ooz006] [Medline: 31984356]
24. Ronquillo JG, Erik Winterholler J, Cwikla K, Szymanski R, Levy C. Health IT, hacking, and cybersecurity: national trends in data breaches of protected health information. JAMIA Open 2018 Jul;1(1):15-19 [FREE Full text] [doi: 10.1093/jamiaopen/ooy019] [Medline: 31984315]
25. Dalal AK, Fuller T, Garabedian P, Ergai A, Balint C, Bates DW, et al. Systems engineering and human factors support of a system of novel EHR-integrated tools to prevent harm in the hospital. J Am Med Inform Assoc 2019 Jun 01;26(6):553-560. [doi: 10.1093/jamia/ocz002] [Medline: 30903660]
26. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1:40 [FREE Full text] [doi: 10.1038/s41746-018-0048-y] [Medline: 31304321]
27. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
28. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science 2019 Feb 22;363(6429):810-812 [FREE Full text] [doi: 10.1126/science.aaw0029] [Medline: 30792287]
29. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. J Am Med Inform Assoc 2014 Oct;21(e2):e304-e311 [FREE Full text] [doi: 10.1136/amiajnl-2013-002316] [Medline: 24674844]
30. Eloff J, Bella M. Software failures: An overview. In: Software Failure Investigation. Cham: Springer; 2018:7-24.
31. Zhou L, Blackley SV, Kowalski L, Doan R, Acker WW, Landman AB, et al. Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists. JAMA Netw Open 2018 Jul;1(3):e180530 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.0530] [Medline: 30370424]
32. Salahuddin L, Ismail Z, Hashim UR, Ismail NH, Raja Ikram RR, Abdul Rahim F, et al. Healthcare practitioner behaviours that influence unsafe use of hospital information systems. Health Informatics J 2020 Mar 07;26(1):420-434. [doi: 10.1177/1460458219833090] [Medline: 30843460]
33. Rodziewicz T, Hipskind J. Medical Error. In: StatPearls [Internet]. Treasure Island: StatPearls Publishing; 2019.
34. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
35. Chai KEK, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. J Am Med Inform Assoc 2013 Sep 01;20(5):980-985 [FREE Full text] [doi: 10.1136/amiajnl-2012-001409] [Medline: 23666777]
36. Awaysheh A, Wilcke J, Elvinger F, Rees L, Fan W, Zimmerman KL. Review of Medical Decision Support and Machine-Learning Methods. Vet Pathol 2019 Jul;56(4):512-525. [doi: 10.1177/0300985819829524] [Medline: 30866728]
37. Sanchez-Morillo D, Fernandez-Granero MA, Leon-Jimenez A. Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: A systematic review. Chron Respir Dis 2016 Aug 23;13(3):264-283 [FREE Full text] [doi: 10.1177/1479972316642365] [Medline: 27097638]
38. Pellegrini E, Ballerini L, Hernandez MDCV, Chappell FM, González-Castro V, Anblagan D, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. Alzheimers Dement (Amst) 2018;10:519-535 [FREE Full text] [doi: 10.1016/j.dadm.2018.07.004] [Medline: 30364671]
39. Safdar S, Zafar S, Zafar N, Khan NF. Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Artif Intell Rev 2017 Mar 25;50(4):597-623. [doi: 10.1007/s10462-017-9552-8]
40. Dallora AL, Eivazzadeh S, Mendes E, Berglund J, Anderberg P. Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. PLoS One 2017;12(6):e0179804 [FREE Full text] [doi: 10.1371/journal.pone.0179804] [Medline: 28662070]

41. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digital Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/s2589-7500(19)30123-2]
42. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. IEEE Access 2017;5:8869-8879. [doi: 10.1109/ACCESS.2017.2694446]
43. Yeh DH, Tam S, Fung K, MacNeil SD, Yoo J, Winquist E, et al. Transoral robotic surgery vs. radiotherapy for management of oropharyngeal squamous cell carcinoma - A systematic review of the literature. Eur J Surg Oncol 2015 Dec;41(12):1603-1614. [doi: 10.1016/j.ejso.2015.09.007] [Medline: 26461255]
44. Ficarra V, Novara G, Rosen RC, Artibani W, Carroll PR, Costello A, et al. Systematic review and meta-analysis of studies reporting urinary continence recovery after robot-assisted radical prostatectomy. Eur Urol 2012 Sep;62(3):405-417. [doi: 10.1016/j.eururo.2012.05.045] [Medline: 22749852]
45. Dowthwaite SA, Franklin JH, Palma DA, Fung K, Yoo J, Nichols AC. The role of transoral robotic surgery in the management of oropharyngeal cancer: a review of the literature. ISRN Oncol 2012;2012:945162 [FREE Full text] [doi: 10.5402/2012/945162] [Medline: 22606380]
46. Karthik K, Colegate-Stone T, Dasgupta P, Tavakkolizadeh A, Sinha J. Robotic surgery in trauma and orthopaedics: a systematic review. Bone Joint J 2015 Mar;97-B(3):292-299. [doi: 10.1302/0301-620X.97B3.35107] [Medline: 25737510]
47. Klock M, Kang H, Gong Y. Scoring Patient Fall Reports Using Quality Rubric and Machine Learning. Stud Health Technol Inform 2019 Aug 21;264:639-643. [doi: 10.3233/SHTI190301] [Medline: 31438002]
48. Wang E, Kang H, Gong Y. Generating a Health Information Technology Event Database from FDA MAUDE Reports. Stud Health Technol Inform 2019 Aug 21;264:883-887. [doi: 10.3233/SHTI190350] [Medline: 31438051]
49. Zhou S, Kang H, Yao B, Gong Y. An automated pipeline for analyzing medication event reports in clinical settings. BMC Med Inform Decis Mak 2018 Dec 07;18(Suppl 5):113 [FREE Full text] [doi: 10.1186/s12911-018-0687-6] [Medline: 30526590]
50. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. JAMA 2019 Jan 01;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
51. Davatzikos C. Machine learning in neuroimaging: Progress and challenges. Neuroimage 2019 Aug 15;197:652-656 [FREE Full text] [doi: 10.1016/j.neuroimage.2018.10.003] [Medline: 30296563]
52. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med 2018 Nov 01;178(11):1544-1547 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
53. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
54. Powers EM, Shiffman RN, Melnick ER, Hickner A, Sharifi M. Efficacy and unintended consequences of hard-stop alerts in electronic health record systems: a systematic review. J Am Med Inform Assoc 2018 Nov 01;25(11):1556-1566 [FREE Full text] [doi: 10.1093/jamia/ocy112] [Medline: 30239810]
55. Choudhury A, Asan O. Patient SafetyArtificial Intelligence. OSF. URL: https://osf.io/vqjk5/ [accessed 2019-09-15]
56. Ecker ED, Skelly AC. Conducting a winning literature search. Evid Based Spine Care J 2010 May;1(1):9-14 [FREE Full text] [doi: 10.1055/s-0028-1100887] [Medline: 23544018]
57. Chen L, Dubrawski A, Wang D, Fiterau M, Guillame-Bert M, Bose E, et al. Using Supervised Machine Learning to Classify Real Alerts and Artifact in Online Multisignal Vital Sign Monitoring Data. Crit Care Med 2016 Jul;44(7):e456-e463 [FREE Full text] [doi: 10.1097/CCM.0000000000001660] [Medline: 26992068]
58. Ansari S, Belle A, Ghanbari H, Salamango M, Najarian K. Suppression of false arrhythmia alarms in the ICU: a machine learning approach. Physiol Meas 2016 Aug;37(8):1186-1203. [doi: 10.1088/0967-3334/37/8/1186] [Medline: 27454017]
59. Zhang Q, Chen X, Fang Z, Zhan Q, Yang T, Xia S. Reducing false arrhythmia alarm rates using robust heart rate estimation and cost-sensitive support vector machines. Physiol Meas 2017 Feb;38(2):259-271. [doi: 10.1088/1361-6579/38/2/259] [Medline: 28099159]
60. Antink CH, Leonhardt S, Walter M. Reducing false alarms in the ICU by quantifying self-similarity of multimodal biosignals. Physiol Meas 2016 Aug;37(8):1233-1252. [doi: 10.1088/0967-3334/37/8/1233] [Medline: 27454256]
61. Eerikäinen LM, Vanschoren J, Rooijakkers MJ, Vullings R, Aarts RM. Reduction of false arrhythmia alarms using signal selection and machine learning. Physiol Meas 2016 Aug 25;37(8):1204-1216. [doi: 10.1088/0967-3334/37/8/1204] [Medline: 27454128]
62. Ménard T, Barmaz Y, Koneswarakantha B, Bowling R, Popko L. Enabling Data-Driven Clinical Quality Assurance: Predicting Adverse Event Reporting in Clinical Trials Using Machine Learning. Drug Saf 2019 Sep 23;42(9):1045-1053 [FREE Full text] [doi: 10.1007/s40264-019-00831-4] [Medline: 31123940]
63. Segal G, Segev A, Brom A, Lifshitz Y, Wasserstrum Y, Zimlichman E. Reducing drug prescription errors and adverse drug events by application of a probabilistic, machine-learning based clinical decision support system in an inpatient setting. J Am Med Inform Assoc 2019 Dec 01;26(12):1560-1565. [doi: 10.1093/jamia/ocz135] [Medline: 31390471]

64. Hu SB, Wong DJL, Correa A, Li N, Deng JC. Prediction of Clinical Deterioration in Hospitalized Adult Patients with Hematologic Malignancies Using a Neural Network Model. PLoS One 2016;11(8):e0161401 [FREE Full text] [doi: 10.1371/journal.pone.0161401] [Medline: 27532679]
65. Kwon J, Lee Y, Lee Y, Lee S, Park J. An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. J Am Heart Assoc 2018 Jun 26;7(13):e008678 [FREE Full text] [doi: 10.1161/JAHA.118.008678] [Medline: 29945914]
66. Gupta J, Patrick J. Automated validation of patient safety clinical incident classification: macro analysis. Stud Health Technol Inform 2013;188:52-57. [Medline: 23823288]
67. Wang Y, Coiera EW, Runciman W, Magrabi F. Automating the Identification of Patient Safety Incident Reports Using Multi-Label Classification. : IOS Press; 2017 Presented at: Precision Healthcare Through Informatics: Proceedings of the 16th World Congress on Medical and Health Informatics; August 21-25, 2017; Hangzhou, China p. 609-613.
68. Fong A, Harriott N, Walters DM, Foley H, Morrissey R, Ratwani RR. Integrating natural language processing expertise with patient safety event review committees to improve the analysis of medication events. Int J Med Inform 2017 Aug;104:120-125. [doi: 10.1016/j.ijmedinf.2017.05.005] [Medline: 28529113]
69. ElMessiry A, Zhang Z, Cooper W, Catron T, Karrass J, Singh M, editors. Leveraging sentiment analysis for classifying patient complaints. 2017 Presented at: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,Health Informatics; 2017; Boston. [doi: 10.1145/3107411.3107421]
70. Chondrogiannis E, Andronikou V, Varvarigou T, Karanastasis E, editors. Semantically-Enabled Context-Aware Abbreviations Expansion in the Clinical Domain. 2017 Presented at: Proceedings of the 9th International Conference on Bioinformatics Biomedical Technology; 2017; Washington DC. [doi: 10.1145/3093293.3093304]
71. Liang C, Gong Y. Automated Classification of Multi-Labeled Patient Safety Reports: A Shift from Quantity to Quality Measure. Stud Health Technol Inform 2017;245:1070-1074. [Medline: 29295266]
72. Ong M, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. J Am Med Inform Assoc 2012 Jun;19(e1):e110-e118 [FREE Full text] [doi: 10.1136/amiajnl-2011-000562] [Medline: 22237865]
73. Taggart M, Chapman WW, Steinberg BA, Ruckel S, Pregenzer-Wenzler A, Du Y, et al. Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. JAMA Netw Open 2018 Oct 05;1(6):e183451 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.3451] [Medline: 30646240]
74. Denecke K, Lutz HS, Pöpel A, May R, editors. Talking to ana: A mobile self-anamnesis application with conversational user interface. 2018 Presented at: Proceedings of the 2018 International Conference on Digital Health; 2018; Lyon. [doi: 10.1145/3194658.3194670]
75. Evans HP, Anastasiou A, Edwards A, Hibbert P, Makeham M, Luz S, et al. Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches. Health Informatics J 2019 Mar 07:1460458219833102. [doi: 10.1177/1460458219833102] [Medline: 30843455]
76. Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. J Am Med Inform Assoc 2019 Dec 01;26(12):1600-1608. [doi: 10.1093/jamia/ocz146] [Medline: 31730700]
77. Li M, Ladner D, Miller S, Classen D. Identifying hospital patient safety problems in real-time with electronic medical record data using an ensemble machine learning model. Int J Clin Med Inform 2018;1(1):43-58.
78. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):848-855. [doi: 10.1001/jama.2011.1204] [Medline: 21862746]
79. Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med Inform Decis Mak 2017 Jun 12;17(1):84 [FREE Full text] [doi: 10.1186/s12911-017-0483-8] [Medline: 28606174]
80. Rosenbaum M, Baron J. Using Machine Learning-Based Multianalyte Delta Checks to Detect Wrong Blood in Tube Errors. Am J Clin Pathol 2018 Oct 24;150(6):555-566. [doi: 10.1093/ajcp/aqy085] [Medline: 30169595]
81. McKnight SD. Semi-supervised classification of patient safety event reports. J Patient Saf 2012 Jun;8(2):60-64. [doi: 10.1097/PTS.0b013e31824ab987] [Medline: 22543364]
82. Marella WM, Sparnon E, Finley E. Screening Electronic Health Record-Related Patient Safety Reports Using Machine Learning. J Patient Saf 2017 Mar;13(1):31-36. [doi: 10.1097/PTS.0000000000000104] [Medline: 24721977]
83. Ye C, Wang O, Liu M, Zheng L, Xia M, Hao S, et al. A Real-Time Early Warning System for Monitoring Inpatient Mortality Risk: Prospective Study Using Electronic Medical Record Data. J Med Internet Res 2019 Jul 05;21(7):e13719 [FREE Full text] [doi: 10.2196/13719] [Medline: 31278734]
84. Fong A, Adams KT, Gaunt MJ, Howe JL, Kellogg KM, Ratwani RM. Identifying health information technology related safety event reports from patient safety event report databases. J Biomed Inform 2018 Oct;86:135-142 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.007] [Medline: 30213556]
85. Simon ACR, Holleman F, Gude WT, Hoekstra JBL, Peute LW, Jaspers MWM, et al. Safety and usability evaluation of a web-based insulin self-titration system for patients with type 2 diabetes mellitus. Artif Intell Med 2013 Sep;59(1):23-31. [doi: 10.1016/j.artmed.2013.04.009] [Medline: 23735522]

86. Song D, Chen Y, Min Q, Sun Q, Ye K, Zhou C, et al. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. J Clin Pharm Ther 2019 Apr 18;44(2):268-275. [doi: 10.1111/jcpt.12786] [Medline: 30565313]
87. Hammann F, Gutmann H, Vogt N, Helma C, Drewe J. Prediction of adverse drug reactions using decision tree modeling. Clin Pharmacol Ther 2010 Jul 10;88(1):52-59. [doi: 10.1038/clpt.2009.248] [Medline: 20220749]
88. Bean DM, Wu H, Iqbal E, Dzahini O, Ibrahim ZM, Broadbent M, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Sci Rep 2017 Nov 27;7(1):16416. [doi: 10.1038/s41598-017-16674-x] [Medline: 29180758]
89. Hu Y, Tai C, Tsai C, Huang M. Improvement of Adequate Digoxin Dosage: An Application of Machine Learning Approach. J Healthc Eng 2018;2018:3948245. [doi: 10.1155/2018/3948245] [Medline: 30210752]
90. Tang Y, Yang J, Ang PS, Dorajoo SR, Foo B, Soh S, et al. Detecting adverse drug reactions in discharge summaries of electronic medical records using Readpeer. Int J Med Inform 2019 Aug;128:62-70. [doi: 10.1016/j.ijmedinf.2019.04.017] [Medline: 31160013]
91. Hu Y, Wu F, Lo C, Tai C. Predicting warfarin dosage from clinical data: a supervised learning approach. Artif Intell Med 2012 Sep;56(1):27-34. [doi: 10.1016/j.artmed.2012.04.001] [Medline: 22537823]
92. Hasan S, Duncan GT, Neill DB, Padman R. Automatic detection of omissions in medication lists. J Am Med Inform Assoc 2011 Jul 01;18(4):449-458 [FREE Full text] [doi: 10.1136/amiajnl-2011-000106] [Medline: 21447497]
93. Labovitz DL, Shafner L, Reyes Gil M, Virmani D, Hanina A. Using Artificial Intelligence to Reduce the Risk of Nonadherence in Patients on Anticoagulation Therapy. Stroke 2017 May;48(5):1416-1419 [FREE Full text] [doi: 10.1161/STROKEAHA.116.016281] [Medline: 28386037]
94. Long J, Yuan MJ, Poonawala R. An Observational Study to Evaluate the Usability and Intent to Adopt an Artificial Intelligence-Powered Medication Reconciliation Tool. Interact J Med Res 2016 May 16;5(2):e14 [FREE Full text] [doi: 10.2196/ijmr.5462] [Medline: 27185210]
95. Reddy M, Pesl P, Xenou M, Toumazou C, Johnston D, Georgiou P, et al. Clinical Safety and Feasibility of the Advanced Bolus Calculator for Type 1 Diabetes Based on Case-Based Reasoning: A 6-Week Nonrandomized Single-Arm Pilot Study. Diabetes Technol Ther 2016 Aug;18(8):487-493. [doi: 10.1089/dia.2015.0413] [Medline: 27196358]
96. Schiff GD, Volk LA, Volodarskaya M, Williams DH, Walsh L, Myers SG, et al. Screening for medication errors using an outlier detection system. J Am Med Inform Assoc 2017 Mar 01;24(2):281-287. [doi: 10.1093/jamia/ocw171] [Medline: 28104826]
97. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015 May 06;15:37 [FREE Full text] [doi: 10.1186/s12911-015-0160-8] [Medline: 25943550]
98. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. Int J Med Inform 2015 Dec;84(12):1057-1064. [doi: 10.1016/j.ijmedinf.2015.09.002] [Medline: 26456569]
99. Tinoco A, Evans RS, Staes CJ, Lloyd JF, Rothschild JM, Haug PJ. Comparison of computerized surveillance and manual chart review for adverse events. J Am Med Inform Assoc 2011;18(4):491-497 [FREE Full text] [doi: 10.1136/amiajnl-2011-000187] [Medline: 21672911]
100. Onay A, Onay M, Abul O. Classification of nervous system withdrawn and approved drugs with ToxPrint features via machine learning strategies. Comput Methods Programs Biomed 2017 Apr;142:9-19. [doi: 10.1016/j.cmpb.2017.02.004] [Medline: 28325450]
101. Cai R, Liu M, Hu Y, Melton BL, Matheny ME, Xu H, et al. Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. Artif Intell Med 2017 Feb;76:7-15 [FREE Full text] [doi: 10.1016/j.artmed.2017.01.004] [Medline: 28363289]
102. Dandala B, Joopudi V, Devarakonda M. Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks. Drug Saf 2019 Jan;42(1):135-146. [doi: 10.1007/s40264-018-0764-x] [Medline: 30649738]
103. Dey S, Luo H, Fokoue A, Hu J, Zhang P. Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinformatics 2018 Dec 28;19(Suppl 21):476 [FREE Full text] [doi: 10.1186/s12859-018-2544-0] [Medline: 30591036]
104. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEx: A System for Detecting Medications, Adverse Drug Events, and Their Relations from Clinical Notes. Drug Saf 2019 Jan;42(1):123-133 [FREE Full text] [doi: 10.1007/s40264-018-0761-0] [Medline: 30600484]
105. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting Adverse Drug Events with Rapidly Trained Classification Models. Drug Saf 2019 Jan 16;42(1):147-156 [FREE Full text] [doi: 10.1007/s40264-018-0763-y] [Medline: 30649737]
106. Lian D, Khoshneshin M, Street WN, Liu M. Adverse drug effect detection. IEEE J Biomed Health Inform 2013 Mar;17(2):305-311. [doi: 10.1109/TITB.2012.2227272] [Medline: 24235108]

107. Huang L, Wu X, Chen JY. Predicting adverse side effects of drugs. BMC Genomics 2011 Dec 23;12 Suppl 5:S11 [FREE Full text] [doi: 10.1186/1471-2164-12-S5-S11] [Medline: 22369493]
108. Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med Inform Decis Mak 2017 Jun 12;17(1):84 [FREE Full text] [doi: 10.1186/s12911-017-0483-8] [Medline: 28606174]
109. Ong M, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. J Am Med Inform Assoc 2012 Jun;19(e1):e110-e118 [FREE Full text] [doi: 10.1136/amiajnl-2011-000562] [Medline: 22237865]
110. Dey S, Luo H, Fokoue A, Hu J, Zhang P. Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinformatics 2018 Dec 28;19(Suppl 21):476 [FREE Full text] [doi: 10.1186/s12859-018-2544-0] [Medline: 30591036]
111. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, Breast Cancer Surveillance Consortium. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015 Nov;175(11):1828-1837 [FREE Full text] [doi: 10.1001/jamainternmed.2015.5231] [Medline: 26414882]
112. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008 Jun 28;336(7659):1475-1482 [FREE Full text] [doi: 10.1136/bmj.39609.449676.25] [Medline: 18573856]
113. Iacobucci G. Computer error may have led to incorrect prescribing of statins to thousands of patients. BMJ 2016 May 13;353:i2742. [doi: 10.1136/bmj.i2742] [Medline: 27178396]
114. NIST. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools. National Institute of Standards and Technology U.S. Department of Commerce. 2019 Sep 09. URL: https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf [accessed 2020-06-20]
115. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
116. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. JAMA Dermatol 2019 Aug 14;155(10):1135. [doi: 10.1001/jamadermatol.2019.1735] [Medline: 31411641]
117. Executive Order 13859 - Maintaining American Leadership in Artificial Intelligence, 84 FR 3967. Federal Register: The Daily Journal of the Unites States Government. Washington DC: The White House; 2019 Feb 14. URL: https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence [accessed 2020-06-20]
118. Statement from FDA Commissioner Scott Gottlieb, M.D. on steps toward a new, tailored review framework for artificial intelligence-based medical devices. US Food and Drug Administration. 2019 Apr 02. URL: https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-steps-toward-new-tailored-review-framework-artificial [accessed 2020-06-20]
119. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
120. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015 May 19;162(10):735. [doi: 10.7326/l15-5093-2]
121. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019 Apr 20;393(10181):1577-1579. [doi: 10.1016/S0140-6736(19)30037-6] [Medline: 31007185]
122. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005 Mar;17(3):299-310. [doi: 10.1109/TKDE.2005.50]
123. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ Digit Med 2020 Mar 9;3(1):30 [FREE Full text] [doi: 10.1038/s41746-020-0229-3] [Medline: 32195365]
124. Paine SJ, Benator SG. JCAHO initiative seeks to improve patient safety. Medscape 2003;15(1):23-24.
125. Yang H, Poly TN, Jack Li YC. Deep into Patient care: An automated deep learning approach for reshaping patient care in clinical setting. Comput Methods Programs Biomed 2019 Jan;168:A1-A2. [doi: 10.1016/j.cmpb.2018.11.007] [Medline: 30527131]
126. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar 12;28(3):231-237 [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
127. Bonafide CP, Localio AR, Holmes JH, Nadkarni VM, Stemler S, MacMurchy M, et al. Video Analysis of Factors Associated With Response Time to Physiologic Monitor Alarms in a Children's Hospital. JAMA Pediatr 2017 Jun 01;171(6):524-531 [FREE Full text] [doi: 10.1001/jamapediatrics.2016.5123] [Medline: 28394995]
128. Winters BD, Cvach MM, Bonafide CP, Hu X, Konkani A, O'Connor MF, Society for Critical Care Medicine AlarmAlert Fatigue Task Force. Technological Distractions (Part 2): A Summary of Approaches to Manage Clinical Alarms With Intent

to Reduce Alarm Fatigue. Crit Care Med 2018 Jan;46(1):130-137. [doi: 10.1097/CCM.0000000000002803] [Medline: 29112077]
129. Hu X. An algorithm strategy for precise patient monitoring in a connected healthcare enterprise. NPJ Digit Med 2019;2:30 [FREE Full text] [doi: 10.1038/s41746-019-0107-z] [Medline: 31304377]
130. Woodward S. Moving towards a safety II approach. J Patient Safe Risk Manage 2019 Jun 08;24(3):96-99. [doi: 10.1177/2516043519855264]
131. Woodward S. Implementing Patient Safety: Addressing Culture, Conditions, and Values to Help People Work Safely. New York: Routledge Productivity Press; 2019.

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
EHR: electronic health record
FDA: Food and Drug Administration
HHS: US Department of Health and Human Services
HIPAA: Health Insurance Portability and Accountability Act
MeSH: Medical Subject Headings
NIST: National Institute of Standards and Technology
PANDIT: Patient Assisting Net-Based Diabetes Insulin Titration
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Edited by C Lovis, G Eysenbach; submitted 06.03.20; peer-reviewed by E Chiou, A Anastasiou, S Pitoglou; comments to author 30.03.20; revised version received 26.05.20; accepted 13.06.20; published 24.07.20.

Please cite as:
Choudhury A, Asan O
Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review
JMIR Med Inform 2020;8(7):e18599
URL: http://medinform.jmir.org/2020/7/e18599/
doi:10.2196/18599
PMID:32706688

Original Paper

Enabling Agile Clinical and Translational Data Warehousing: Platform Development and Evaluation

Helmut Spengler[1], Dipl Inf; Claudia Lang[1], MSc; Tanmaya Mahapatra[1], PhD; Ingrid Gatz[1], MSc; Klaus A Kuhn[1], MD, PhD; Fabian Prasser[2,3], PhD

[1]Institute of Medical Informatics, Statistics and Epidemiology, University Medical Center rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany

[2]Charité - Universitätsmedizin Berlin, Berlin, Germany

[3]Berlin Institute of Health, Berlin, Germany

Corresponding Author:
Fabian Prasser, PhD
Charité - Universitätsmedizin Berlin
Charitéplatz 1
Berlin
Germany
Phone: 49 30450 ext 528781
Email: fabian.prasser@charite.de

Abstract

Background: Modern data-driven medical research provides new insights into the development and course of diseases and enables novel methods of clinical decision support. Clinical and translational data warehouses, such as Informatics for Integrating Biology and the Bedside (i2b2) and tranSMART, are important infrastructure components that provide users with unified access to the large heterogeneous data sets needed to realize this and support use cases such as cohort selection, hypothesis generation, and ad hoc data analysis.

Objective: Often, different warehousing platforms are needed to support different use cases and different types of data. Moreover, to achieve an optimal data representation within the target systems, specific domain knowledge is needed when designing data-loading processes. Consequently, informaticians need to work closely with clinicians and researchers in short iterations. This is a challenging task as installing and maintaining warehousing platforms can be complex and time consuming. Furthermore, data loading typically requires significant effort in terms of data preprocessing, cleansing, and restructuring. The platform described in this study aims to address these challenges.

Methods: We formulated system requirements to achieve agility in terms of platform management and data loading. The derived system architecture includes a cloud infrastructure with unified management interfaces for multiple warehouse platforms and a data-loading pipeline with a declarative configuration paradigm and meta-loading approach. The latter compiles data and configuration files into forms required by existing loading tools, thereby automating a wide range of data restructuring and cleansing tasks. We demonstrated the fulfillment of the requirements and the originality of our approach by an experimental evaluation and a comparison with previous work.

Results: The platform supports both i2b2 and tranSMART with built-in security. Our experiments showed that the loading pipeline accepts input data that cannot be loaded with existing tools without preprocessing. Moreover, it lowered efforts significantly, reducing the size of configuration files required by factors of up to 22 for tranSMART and 1135 for i2b2. The time required to perform the compilation process was roughly equivalent to the time required for actual data loading. Comparison with other tools showed that our solution was the only tool fulfilling all requirements.

Conclusions: Our platform significantly reduces the efforts required for managing clinical and translational warehouses and for loading data in various formats and structures, such as complex entity-attribute-value structures often found in laboratory data. Moreover, it facilitates the iterative refinement of data representations in the target platforms, as the required configuration files are very compact. The quantitative measurements presented are consistent with our experiences of significantly reduced efforts for building warehousing platforms in close cooperation with medical researchers. Both the cloud-based hosting infrastructure and the data-loading pipeline are available to the community as open source software with comprehensive documentation.

(JMIR Med Inform 2020;8(7):e15918) doi:10.2196/15918

KEYWORDS

cohort selection; hypothesis generation; data warehouse; translational research; hosting; Docker; extract-transform-load; i2b2; tranSMART

Introduction

Background

Digitalization of health care promises to enable personalized and predictive medicine [1]. On the basis of digital data that characterize patients and probands at comprehensive depth and breadth [2], modern methods of data analytics can be used to detect unknown relationships between biomedical parameters, discover new patterns, and enable decision support systems by using this knowledge to infer or predict parameters, for example, diagnoses or outcomes [3,4]. A *learning health system* [5], which makes health care data available for secondary research purposes, is an important building block of this development. By comprehensive data integration within and across sites, a massive change in clinical and research processes is envisioned, which will accelerate translation and lead to measurable benefits for patients [6]. In this study, we focus on the integration of structured, that is, typically tabular, clinical and research data.

Multiple technical challenges must be addressed to provide the large, high-quality data sets needed for such purposes. Data from distributed and heterogeneous sources must be integrated at the technical, structural, and semantic levels [7]. To this end, a 3-step extraction-transformation-loading (ETL) process is often implemented:

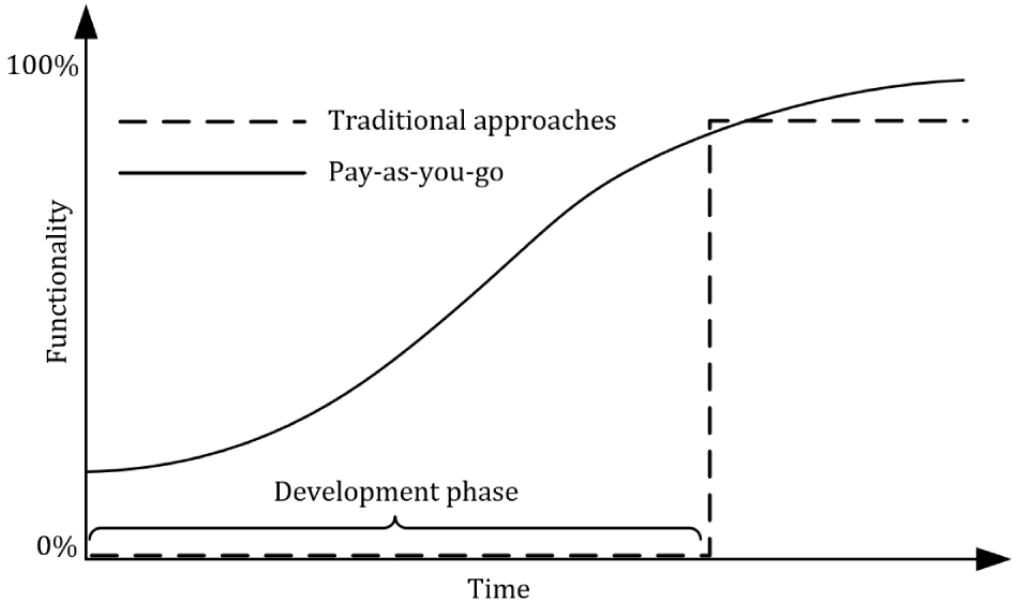
1. Data from research and health care systems are transferred into a staging area in the form of nearly exact copies of data extracted from the source systems [8].
2. Within the staging area, the structure, syntax, and semantics of these data extracts are then normalized into a common data model (CDM) using standard terminologies. These common data representations typically implement a specific database schema, which efficiently and effectively supports complex analytical query processing.
3. Finally, the data are loaded into the target system.

Important examples include clinical and translational data warehousing platforms, such as Informatics for Integrating Biology and the Bedside (i2b2) [9], tranSMART [10], and the Observational Medical Outcomes Partnership (OMOP) CDM [11]; federated and distributed solutions, such as the Shared Health Research Information Network [12]; and the tools provided by Observational Health Data Sciences and Informatics (OHDSI) [11], which can be deployed on top of these analytical databases.

These existing biomedical data analytics platforms offer a wide range of functionalities and integrate different software solutions for data storage, workflow orchestration, and data analysis using multi-tier architectures. As a result of this complexity, considerable technical expertise is required to set them up in a secure manner. These challenges increase even further when organizations run several data-driven research projects and hence need to set up, configure, and maintain multiple warehouse instances. Moreover, ensuring that input data are represented in the analytics platforms in a sound structure with reasonable semantics requires significant medical expertise. It is well known that bridging the interdisciplinary gap between these two worlds requires iterative development processes, in which different solutions are evaluated in short feedback cycles [13]. As we will show later, existing data-loading tools for the aforementioned platforms, however, typically require complex configuration files and input data that adhere to specific formats and structures. Consequently, substantial data restructuring and cleansing is required before data loading can be started and initial feedback can be collected.

In an ideal world, upfront efforts for project-specific technical setup, data cleansing, and data structuring can be avoided, and development starts rapidly, while repeated discussions with clinicians and medical researchers are carried out in parallel [14]. Technical solutions that facilitate this approach have been called *dataspace management systems* [15]. The key idea is to implement a *pay-as-you-go* approach to data integration. A comparison with traditional approaches is presented in Figure 1. It illustrates how the traditional approaches are characterized by an initial development phase in which the data are being integrated on a syntactic, structural, and semantic level, and no service is provided to the users. In contrast, the pay-as-you-go approach provides some initial functionality from the beginning, which is then incrementally extended to better meet the requirements [15,16]. This means that the associated development process can be carried out in an agile manner, involving close cooperation and short feedback cycles with end users. This comes with multiple benefits for the parties involved: clinicians or medical researchers are provided with initial functionalities much more quickly, and feedback can be provided to the development team more often. This is particularly important for data loading because it has been estimated that the development of ETL processes accounts for up to 70% of the total effort required to set up data warehouses [7,17]. For both end users and developers, this can also lead to the reduction of duplicate and redundant work, thus significantly reducing the efforts required. The approach is related to agile methods of software engineering, in which software evolves through continuous collaboration between developers and users. It is well known that this can also help to better bridge the interdisciplinary gaps [18].

Figure 1. Schematic comparison of traditional approaches to data integration and the pay-as-you-go approach.



Objectives

The aim of this study was to implement a platform that enables the deployment and customization of well-known clinical and translational data warehousing solutions in close cooperation with end users in an agile approach. Our solution consists of 2 parts with the following unique features:

1. A cloud-based warehouse management infrastructure, which supports the installation and maintenance of i2b2 and tranSMART in an integrated manner by providing a common set of commands; implements security-by-default features, including transport layer encryption, host-based access control, and password management; and is based on verifiable and authenticatable software to enable installations within high-security perimeters of hospital information technology (IT) environments.
2. A flexible data-loading pipeline, which supports loading data into both i2b2 and tranSMART; is able to process heterogeneous data with different degrees of structure and cleanliness; and performs automated data cleansing and preprocessing, including automatic detection of the syntax and format of input data, and has the ability to handle different encodings as well as missing and duplicate data.

The complete software stack is available to the community as open source software [19,20]. In this study, we provide readers with an overview of the most important system requirements and design decisions. To demonstrate that our solution enables an agile approach to be implemented in a professional context, we present the results of a structured comparison with existing management infrastructures and data-loading pipelines as well as an experimental evaluation of data-loading processes. Our results show that our management infrastructure is the only publicly available open source implementation that supports all the abovementioned features, which is essential for secure deployments in professional IT environments. Moreover, the experimental evaluation showed that no other open source data-loading pipeline was able to process 3 different benchmark data sets, including structured research data, complex longitudinal clinical data, and highly structured billing data, in their raw form. The experiments also showed that our solution is feasible from a computational perspective. We believe that the software presented in this study can be an important tool to support medical informaticians with realizing data warehousing projects and that the methods implemented can provide system developers with novel ideas for the development of future platforms.

Methods

Selection of Target Systems

Clinical and translational data warehouses provide users with efficient analytical access to integrated data sets [21,22]. As an initial step, we decided to utilize an infrastructure supporting i2b2 and tranSMART as both of these have a broad installed base and strong community support. For example, the integrated solution of Hôpital Européen Georges-Pompidou [23] uses i2b2 and tranSMART, integrating data from electronic patient records, including aggregated, anonymized, and *deanonymized* patient data. The tranSMART platform [10] is based on the i2b2 framework, and its suitability for data from clinical studies has already been demonstrated in various projects [24]. In combination, they can be used to support a wide range of use cases.

The i2b2 platform is very well suited for representing longitudinal and often semistructured clinical data, and it supports complex features such as temporal queries against time series data [9]. TranSMART was built over the i2b2 data model to provide improved support for high-dimensional data. The system is well suited for integrating structured research data as well as high-throughput data, and it provides comprehensive support for ad hoc graphical data analysis and cohort comparison [10]. TranSMART offers built-in support for various types of omics data, such as protein and gene expression arrays, single-nucleotide polymorphism data, and certain types of genomic variants. With the recent merger of the i2b2 Foundation and the tranSMART Foundation, a process has been started to

unify both platforms. Until a combined solution becomes available, installations of both systems are needed to support different use cases and to handle different types of data.

The 2 systems offer web-based graphical user interfaces. TranSMART employs a classical three-tier information system architecture, whereas i2b2 consists of an extendable framework consisting of several *cells*. Both platforms can be installed on top of different database management systems. As we focus on open source software, we decided to use PostgreSQL, an open source relational database management system.

Cloud Infrastructure for Managing i2b2 and tranSMART

Rationale and Requirements

Both i2b2 and tranSMART offer a wide range of functionalities, and they are based on a software architecture that integrates components for data storage, workflow orchestration, and data analysis. Consequently, installation, configuration, and maintenance procedures are complex and require solid technical expertise. Concurrently, documentation is often lacking. As an example, the number of tranSMART software dependencies is very large, which regularly leads to some dependencies not being up to date or having become incompatible with the underlying (operating) system infrastructure, requiring manual changes to installation scripts. In contrast, the i2b2 installation process is fairly robust, well documented, and up to date [25]. However, it can be quite challenging to debug configuration errors of i2b2 owing to its highly modular architecture, which involves exchange of complex data via web services. These challenges increase significantly when a larger number of instances need to be set up, configured, and maintained. Furthermore, when deploying such systems in production environments, additional aspects such as transport encryption and password management need to be considered. These and further functionalities are not supported by existing cloud-based deployment solutions for i2b2 and tranSMART, such as the Integrated Data Repository Toolkit (IDRT) [26], i2b2 Quickstart [27], or the prebuilt images available on Docker Hub [28] (see the *Discussion* section for an in-depth comparison).

We, therefore, decided to employ clean virtual containers, ideally together with associated maintenance scripts to quickly boot up, configure, and shut down instances of i2b2 and tranSMART in a uniform manner. The most important requirements were as follows:

1. *Robust installation of a trusted runtime environment:* The solution developed shall streamline the complex installation process of tranSMART and enable rapid instantiation of new instances of tranSMART and i2b2.
2. *Unified installation and maintenance:* The solution shall provide a façade encapsulating important configuration options and make the effective management of multiple instances of i2b2 and tranSMART straightforward by providing easy-to-use common commands for both platforms.
3. *Built-in security:* The solution shall significantly improve the security of i2b2 and tranSMART by enabling transport encryption and host-based access control by default as well as by automatically setting nontrivial passwords.

Technical Design

The cloud infrastructure has been designed to run on a physical or virtual machine with a standard Linux operating system. In this system, Docker needs to be installed as a virtualization platform that enables the provisioning of software in deployment units called containers. Each container encapsulates a complete software stack together with all required dependencies, such as libraries and configuration files. Docker employs OS-level virtualization, which means that in contrast to full virtualization, where each virtual machine contains and runs its own operation system, Docker containers can share one single operating system instance and are thus more lightweight than virtual machines. Although containers are isolated from each other, they can be enabled to communicate through definable channels (eg, Transmission Control Protocol ports). Containers can quickly be instantiated and customized via runtime parameters in this process.

We chose Docker for the following reasons: (1) it enables describing and documenting installation processes in a machine and human-readable format, thus fulfilling our requirement for robust installation and quick instantiation; (2) it allows customizing running containers by means of runtime parameters (eg, access permissions, passwords, and instance names), thus fulfilling our requirement to provide uniform configuration and maintenance scripts for both platforms; (3) its efficient use of resources allows rapid booting up and shutting down instances; and (4) it integrates well with common software development infrastructures, such as GitLab.

As a gateway component to provide transport encryption, host-based access control, and data routing for the particular warehouse instances, we decided to include the Apache HTTP Server into the host environment utilizing its proxy and virtual host modules.

Meeting Requirement 1: Robust Installation of a Trusted Runtime Environment

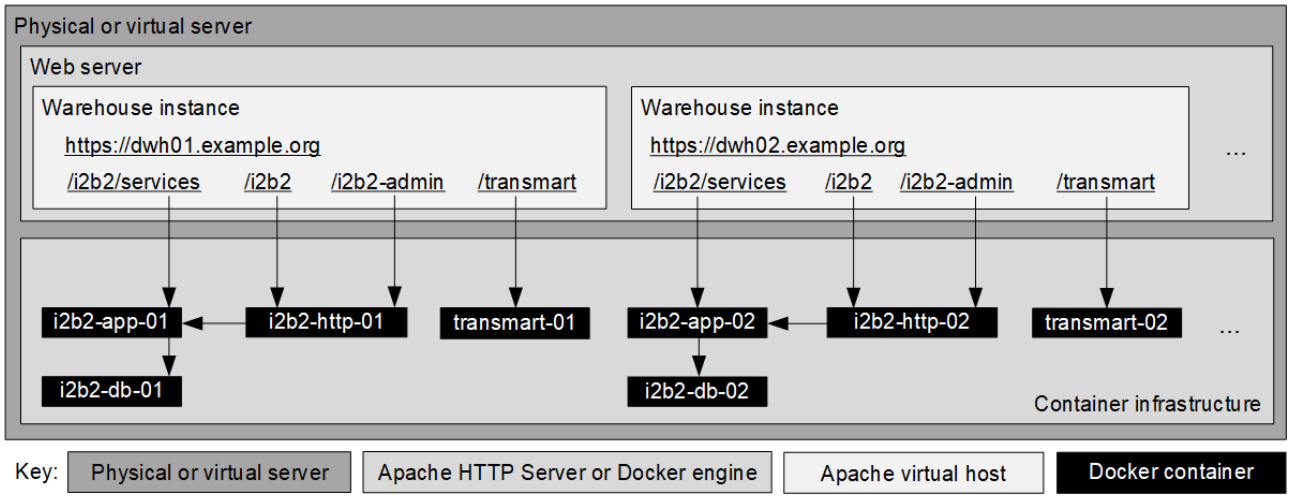
The solution can be used to host an arbitrary number of i2b2 and tranSMART instances. Each host system includes the following containers per instance: (1) a database server for i2b2, (2) an application server for i2b2, (3) a web server for i2b2, and (4) a complete tranSMART software stack. It can be accessed via specific URLs. The subdomain in this URL denotes the warehouse instance, for example, *dwh01* or *dwh02*. Each subdomain is represented by a dedicated Apache virtual host and provides one instance of i2b2 and one instance of tranSMART. As an example, the URL-pattern *[http|https]://dwh02.example.org/i2b2/* denotes the web front-end of i2b2 instance *02*, which is exposed by the Apache virtual host *dwh02.example.org*.

Both tranSMART and i2b2 expose specific ports to provide specific services. These include their web front-ends and various web services. To avoid port clashes when running multiple warehouse instances and their respective containers, the ports used by each container are mapped to corresponding ports on

the host system using specific offsets such that a certain set of ports uniquely identifies each service of each container.

Figure 2 illustrates the components used by the environment and their interactions. The actual instances of i2b2 and tranSMART are implemented as (stacks of) Docker containers (black boxes). Access to these containers is relayed by an Apache web server, which acts as a gateway. Each warehouse instance is represented by a virtual host of the gateway and is identified by the first part of the hostname contained in the URL of the request. Detailed installation instructions along with well-documented configuration files are available on the web [19].

Figure 2. Schematic overview of the components for the provisioning of multiple warehouse instances and their interaction.



Meeting Requirement 2: Unified Installation and Maintenance

To support unified management for instances of both types of systems, we have developed 2 configuration scripts that can be parameterized. Target instances are identified by their type and consecutive numbers (eg, i2b2-04). The first script can be used to set up new warehouse instances and to reset existing instances. It does so by creating configurations for Apache's proxy and virtual host modules and environment files for the Docker compose scripts. If needed, the resulting files can be edited by the administrator (eg, to replace randomly generated passwords) before the new instances are created. The second script can be used for starting, stopping, and deleting warehouse instances as well as associated disk volumes. It has been implemented as a wrapper for Docker compose commands that access the environment variables defined in the associated environment files.

Meeting Requirement 3: Built-In Security

The setup process implements several crucial security measures, including transport layer encryption, server authentication, restricted access paths, and nontrivial default passwords.

Access to the services running on each server is only permitted indirectly via the Apache HTTP Server, which acts as a central gateway. This component takes care of the transport encryption and server authentication mentioned above as well as address-based access control. The only service that can be reached without having to pass the gateway is the database system to enable efficient data loading. Here, access control is implemented at the database level. Permission to access the database has to be granted explicitly, which includes the declaration of address ranges with specific access rights. To simplify the Transport Layer Security configuration, we make use of the *subject alternative name* extension to the X.509 server certificates [29], which our platform uses for authenticating the data warehouses and for transport layer encryption. Embedded plain text secrets and the fact that the source and content of many images cannot be verified have been identified as major risks for system components based on container technologies [30]. This impedes the use of prebuilt images in high-security IT environments. Our infrastructure does not suffer from these shortcomings as we employ Docker Content Trust [31] to verify the authenticity of all base images used. As the current images for i2b2 and tranSMART do not support this authentication mechanism, we decided to build our own images based on authenticated sources (by verifying Pretty Good Privacy signatures of binaries used and/or building them from source). Secure default passwords are automatically created via a random password generator [32] with a default length of 10 characters and injected into the containers at runtime.

Generic and Agile Data-Loading Pipeline for i2b2 and tranSMART

Rationale and Requirements

Populating i2b2 and tranSMART with data is cumbersome and requires significant expertise regarding the underlying database schema and how both systems use it. For this reason, several tools have been developed to simplify this process, including tranSMART-ETL [33], tMDataLoader [34], transmart-batch [35], Integrated Curation Environment (ICE) [36], IDRT [26], transmart-copy [37], and TranSMART data curation toolkit (tmtk) [38]. However, none of these tools fulfill the requirements needed to implement agility (see the *Discussion* section for an in-depth comparison).

First, all available data loaders except transmart-batch are strongly tied to 1 of the 2 target systems. As both are often needed in parallel, this introduces additional preprocessing and configuration efforts. The main reason is that loaders for different systems make different assumptions about the degree of structure and cleanliness of import data. In addition, different loaders use different configuration mechanisms. Moreover, existing tools follow imperative configuration paradigms, where it must be specified how the loading process should be executed, making this process complex and requiring substantial technical expertise as well as domain knowledge. Finally, to support agile and fast loading, tools should be able to automatically handle heterogeneity and errors in input data, such as differences in data encoding and syntax as well as missing and duplicate data. To address these challenges, we needed a data-loading pipeline fulfilling the following requirements:

1. *Platform independence:* The data-loading pipeline shall be designed independent of a specific target system, enabling the loading of data into both i2b2 and tranSMART with the same pipeline using the same configuration files.
2. *Support for different types of data:* The pipeline shall support heterogeneous data with different degrees of structure and cleanliness, such as structured research data, complex longitudinal clinical data, and highly structured billing data, without requiring complex preprocessing or configuration efforts.
3. *Automated data cleansing and preprocessing:* The pipeline shall automatically detect the syntax and format of input data and handle different encodings as well as missing and duplicate data. This significantly reduces efforts and improves agility when providing warehousing solutions.

Technical Design

The most important design decision made to fulfill the requirements listed above was to center the tool around a declarative and model-driven way of configuring the import process. The basic idea was to enable users to match data to an entity-relationship (ER) model that describes the desired target representation of the data. The tool then automatically determines how the input data must be interpreted, transformed, and loaded to reflect this model in the target database. This includes the automatic creation of the ontologies required by i2b2 based on this model. This is in stark contrast to the imperative configuration paradigm found in most ETL tools for i2b2 and tranSMART and significantly reduces the complexity of configuration files and hence efforts (see the *Results* and *Discussion* sections). Moreover, the approach enables our tool to automatically perform a wide range of data transformation and cleansing tasks, thus fulfilling our requirements to support different types of data and automate data cleansing. To fulfill the requirement of platform independence, our tool acts as a *compiler* for configuration files to be used for different ETL tools for i2b2 and tranSMART.

The data-loading tool has been developed in Java using the Spring Batch framework for robust, maintainable, and extensible orchestration of the individual steps of the ETL process; the Univocity parser for reading and writing comma-separated values (CSV) files; and juniversalchardet, a Java port of Mozilla's library, for the automatic detection of file encodings. Access to the target relational database systems has been implemented using Java Database Connectivity.

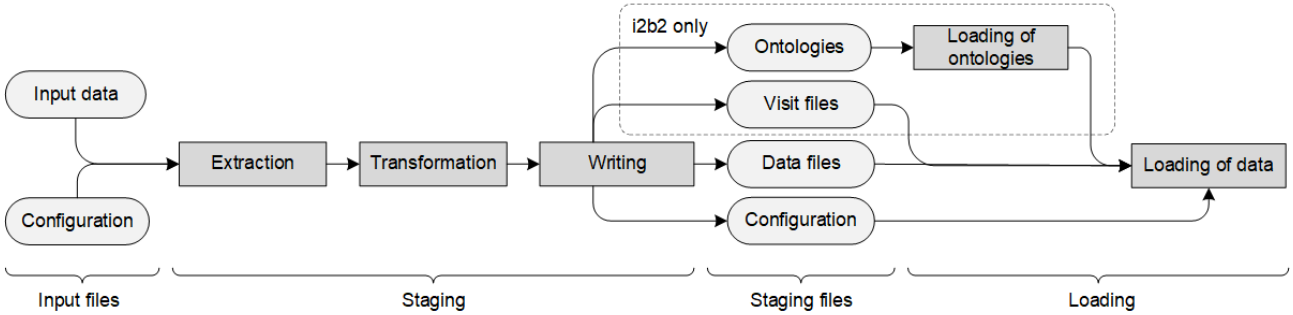
Meeting Requirement 1: Platform Independence

As some powerful loading tools for the different target platforms have already been developed, we decided to implement a meta-loading process consisting of 2 phases: the first is the *staging* phase, in which data are transformed into an intermediate staging representation and configuration files are compiled into the target configuration language for the respective loading tool, which we term *back-end* loader in the context of our meta-loading process. We refer to the transformed data and the configuration files created in this phase as s*taging files*. The second is the *loading* phase, in which the staging files are used to execute the respective back-end loader for the chosen target platform.

Figure 3 illustrates a typical staging and loading process. The *staging phase* is divided into 3 subphases: data extraction, data transformation, and data writing. In the data extraction subphase, our tool reads the declarative configuration, which describes the structure of data to be represented in the target system. On the basis of this configuration, it reads and parses the input data. Details are presented in the 2 subsequent sections. In the data transformation subphase, different data cleansing steps are performed, which are also be presented in the 2 subsequent sections. The last subphase involves writing the transformed data into intermediate files, which are consumed by the back-end data loaders in the loading phase. In the case of i2b2, visit data are written separately. This is followed by writing the associated configuration files, describing how the staging data are to be loaded. In the case of i2b2, this (pre-)final step is concluded by writing data describing the underlying ontologies into separate files. In the *loading phase*, the actual data loading is performed by executing the user-defined back-end loaders. If i2b2 has been selected as the target system, this step is preceded by loading the ontology trees into the target system. Currently, our tool supports the following 2 back-ends for data loading:

- *tMDataLoader*, which has been implemented in Groovy and in stored procedures of the underlying database system to automate data loading for tranSMART [34]. The tool relies on a specific directory structure, containing the data sets and configurations, thus following the convention over configuration approach. It supports the full spectrum of features provided by tranSMART, including the annotation of selected values with timestamps.
- *transmart-batch*, which is implemented in Groovy using Spring Batch and which has been designed to support both tranSMART and i2b2. It requires a specific set of files to be provided about subjects and visits as well as further files containing the actual payload data. It supports fewer features of tranSMART than tMDataLoader and requires significant data cleansing to be performed upfront to provide data in the syntax and structure required.

Figure 3. Overview of data staging and loading with the tool developed. i2b2: Informatics for Integrating Biology and the Bedside.



Meeting Requirement 2: Support for Different Types of Data

As mentioned before, the configuration is performed using a *declarative* approach [39]. This means that users do not need to specify how data should be loaded, but instead map an ER model to the data files to describe the relationship between input and output data. Consequently, the tool can perform a wide range of data transformations automatically without prior normalization, including the automatic creation of the target ontology. Although users are less flexible in defining how data should be represented in the target system, a decent representation can typically be achieved for almost all of the data items, as we will show later, with just a fraction of the effort required to use a more versatile loader. If needed, users can still modify and fine-tune the intermediate staging files to achieve an optimal representation.

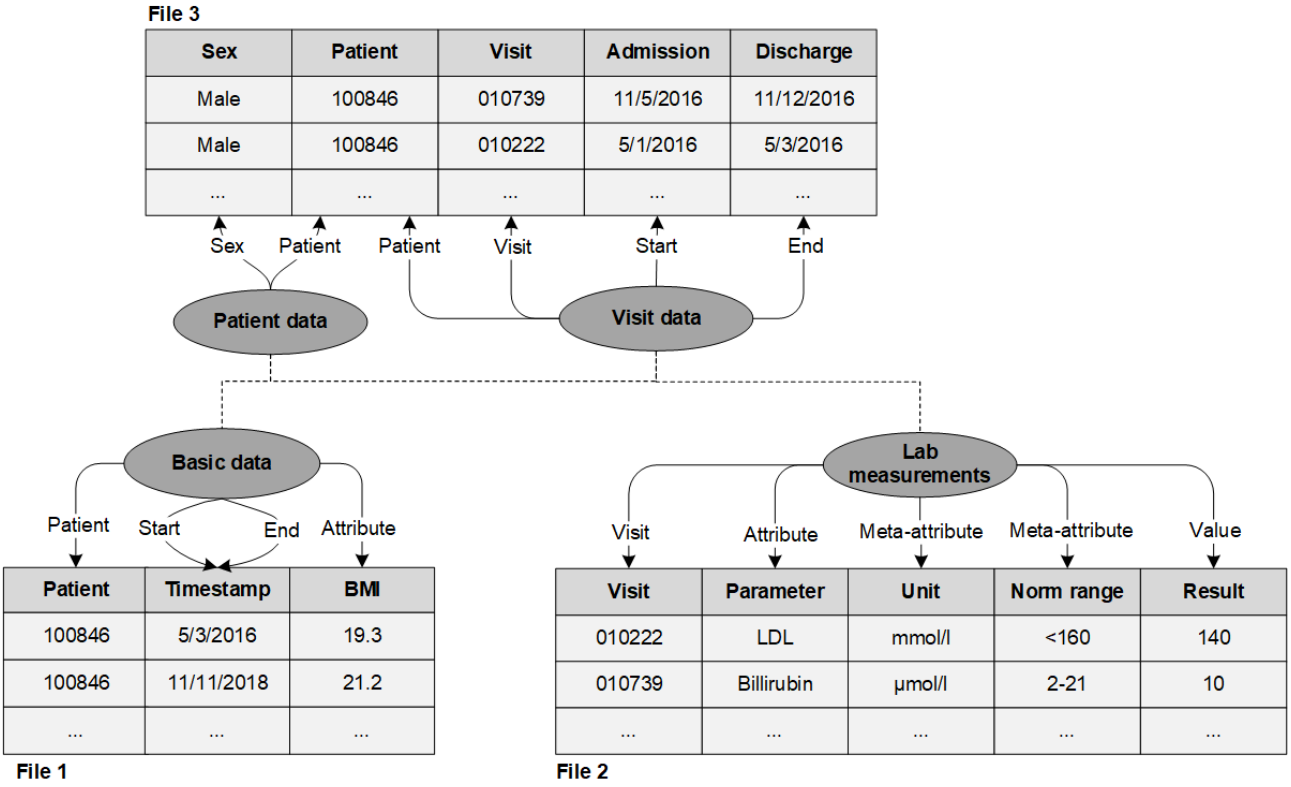
The tool developed was designed in such a way that the maximum degree of the work that needs to be done for successful loading is automated. There are just a few assumptions that are made about input data: (1) data must be tabular, as this is in our experience the most typical format in which clinical and research data can be provided; (2) every line within a file must contain data for a specific patient, visit, or encounter; (3) patients, visits, or encounters must be identified by (composite) keys or timestamps; (4) one file must contain information about the patients or probands—a file describing visits or encounters is optional; and (5) entities may be related to patients, visits, or encounters. Providing information on time points is optional but recommended.

Figure 4 provides an example of how the tool is configured. As can be seen, users are able to specify *entities* that are related to a certain patient or visit and that have *attributes*. Attributes can be mapped to specific columns in the input files. Attributes can be annotated with *meta-attributes*, which are attributes that further specify a specific value for an attribute of a specific entity. In i2b2, these are mapped to *modifiers*. Although there is no direct support for meta-attributes in tranSMART, they can in some cases be represented by creating multiple variants of an attribute that encodes the values of the associated meta-attributes. In addition, there are specific attributes for specifying timestamps and patient or visit identifiers.

The figure also shows an example of how data stored in an *entity-attribute-value* (EAV) model can automatically be denormalized. The EAV model is often used in data collection systems when a large number of different observations are recorded but only a few of them typically apply to a specific patient or proband (eg, lab values). To support this, an additional property *value* is introduced, which can be used to specify how data in EAV form should be denormalized. In the example, one entity will be created in the target systems for each instance of the column *Parameter* having the value from the column *Result* and being annotated with meta-attributes *Unit* and *Norm range*. This is implemented by parsing the input files and populating the configuration with automatically generated parameters for each EAV-encoded data item.

By specifying basic patient, visit, and observational data, the specified EAV entities, the patient data, the observations, an internal model of the ontology, and optionally the associated visits are automatically created. Furthermore, by mapping patients to visits and by relating entities to visits or patients, implicit relationships between the different types of data are constructed. These will also be reflected within the target systems.

Figure 4. Simplified example of an annotation of input files with entities, attributes and relationships. LDL: low density lipoprotein.



Meeting Requirement 3: Automated Data Cleansing and Preprocessing

There are multiple additional features that have been added to the tool based on our experiences with loading a wide range of real-world data sets, which help enforce the syntactic and structural integrity of the input data and which are particularly important due to the heterogeneity of the data sources with respect to these parameters. Important examples include the automated detection of charsets and syntax of input data as well as the automated detection of data types of variables. Features that help enforce semantic integrity include the detection and handling of duplicate data, inconsistent timestamps, and missing values. Finally, support for data filtering and methods for handling uncertainty in timestamps are provided. On a technical level, these tasks are executed as part of either the data loading or the data transformation subphase.

Experimental Design

We evaluated our solution by performing an experimental evaluation of our data-loading approach using different real-world data sets. In the experimental evaluation, we focused on 3 different aspects:

1. *Flexibility:* To demonstrate that our loading tool is able to perform automated data cleansing and restructuring, we used it to load three different types of data sets with varying degrees of structure and cleanliness. Moreover, we also tried to load these data sets using existing data-loading tools to demonstrate that they are not able to process them without prior data cleansing.
2. *Reduced efforts:* To demonstrate that the declarative configuration paradigm of our loading tool significantly reduces the effort required, we compared the number of lines in the configuration files for our tool with the number of lines of the configuration files generated for and needed by existing data-loading tools.
3. *Scalability:* To demonstrate that our approach is computationally feasible, we compared the time needed for automated data cleansing and preprocessing with the time required for actual data loading.

In the experiments, we used real-world data sets from 3 different previous projects: (1) a research data set including *microbiome profiles*, (2) clinical data on *multiple sclerosis*, and (3) *billing data*.

The microbiome profile data set was collected in a study context by our internal medicine department in 2019 and included general information about the probands, lifestyle information obtained through questionnaires, and microbiome profiles (species identified by 16S rRNA gene sequencing) generated from sampled stool, feces, and esophagus tissue. The multiple sclerosis data set was collected by our neurology department since 2010 in the health care context and consisted of longitudinal clinical data, including diagnoses, procedures, clinical scores, medication, lab values, references to biosamples, and metadata of imaging tests. The billing data set consisted of discharge data collected in our hospital in the years 2015-2017 containing demographics and visit data including ventilation time, diagnoses, and procedures. Further details on the projects and use cases supported by these data sets are presented in the *Discussion* section.

For loading data into i2b2, we used the transmart-batch backend, and for loading data into tranSMART, we used the tMDataLoader backend of the pipeline. The experiments were performed with the warehouse instances hosted on a server with Intel Xeon central processing units (CPUs) running at 2.4 GHz with 80 cores, along with 512 GB RAM and 16 TB hard-drives using kernel-based virtual machines provided by Quick EMUlator 2.5.0 running on Ubuntu 18.04. The ETL processes were executed on a desktop machine equipped with a quad-core 3.2 GHz Intel Core i5 CPU running a 64-bit Windows NT kernel, with a 32-bit Java Virtual Machine (1.8.0_202_x86), and with the data input files located on the local file system.

Results

Experiment 1: Flexibility of the Loading Process

In this section, we present results on the flexibility of the loading process for our evaluation data sets and both i2b2 and tranSMART as target systems. The basic properties of the data sets and their representations in the target systems are shown in Table 1.

The microbiome data set originates from a study context and is highly structured. For this reason, and as can be seen in Table 1, i2b2 and tranSMART were both fully able to represent the data set as is. The multiple sclerosis data set, in contrast, was collected in the health care context and consisted of longitudinal clinical data with less structure and a multitude of detailed measurements, such as laboratory values. As can be seen in Table 1, tranSMART could only capture parts of these data (fewer facts by a factor of 6 compared with i2b2) because of missing support for complex time series data and meta-attributes. The billing data set was also highly structured and contained dates of admission and discharge as well as coded diagnoses and procedures. In general, these data could be represented well in i2b2 as well as tranSMART, but the latter system was not able to capture meta-attributes, for example, of diagnoses, resulting in some loss of information.

We emphasize that loading into the different target systems was achieved using the same configuration files. We conclude that our tool provides a high degree of flexibility but that the different target systems are not able to capture all aspects of input data. In general, i2b2 is more suited for representing longitudinal clinical data, and tranSMART is better suited for analyzing highly structured research data.

We further emphasize that our loading pipeline was the only tool with which we were able to load all the data sets described in their raw form without prior transformations or preprocessing. In the remainder of this section, we will briefly cover the issues encountered when using existing open source loading software. We present a detailed comparison with our approach in the *Discussion* section.

When loading the data sets into i2b2, we encountered the following issues: transmart-batch for i2b2 requires the extraction and loading of concept trees into i2b2 before the import of the actual facts. This process is not supported by the tool, and import files also need to be annotated with codes associated with the ontology nodes in the database in an additional preprocessing step. The loading pipeline of IDRT is no longer maintained (over 2.5 years old) and is not compatible with i2b2 1.7.09c and higher, resulting in various errors during data loading. When loading the data sets into tranSMART, we noticed the following problems: tMDataLoader, tmtk, transmart-batch, and ICE could not load the clinical data set where multiple values were provided for the same variable and subject in the same visit. Furthermore, values are required to conform to predefined formats (eg, “yyyy-mm-dd hh:mm:ss” for dates), requiring preprocessing. Transmart-copy could not load any of the data sets used in our experiments without significant preprocessing at the structural and syntactical level, as it required input data to precisely conform to the target schema. TranSMART-ETL could also not load the clinical data set as it was not able to handle missing values. Moreover, it required specific column separators and number formats to be used, requiring input files to be preprocessed accordingly.

Table 1. Overview of the properties of the data sets used in the projects.

Data set	Microbiome profiles	Multiple sclerosis	Billing data
Number of input files	15	19	11
Size of input files in MB	1	497	252
Patients	~50	~7000	~100,000
Visits	~100	~40,000	~300,000
Facts in i2b2[a]	~90,000	~4,600,000	~6,200,000
Facts in tranSMART	~90,000	~750,000	~3,800,000

[a]i2b2: Informatics for Integrating Biology and the Bedside.

Experiment 2: Reduction of Efforts

In this section, we present the results of the reduction of efforts that can be achieved by using our loading tool. We captured this aspect by analyzing the size of files used for actual data loading, which are shown in Table 2. It shows the complexity of configuration files required for data loading with our tool compared with the complexity of the configuration files generated for the backing data loaders. As can be seen, the tool presented in this study generated a large number of files for the different specified entities. Moreover, as a result of the automated denormalization of EAV data and the automated detection of data types, configuring data loading with our tool required significantly fewer lines of configuration parameters

than what would have been required using transmart-batch or tMDataLoader. The configuration files for tranSMART for the multiple sclerosis and the billing data sets were much smaller than the corresponding files for i2b2, as they did not include specifications for meta-attributes.

For the microbiome data set, configuration files for our tool were smaller by factors of between 17.7 (i2b2) and 22.1 (tranSMART). For the multiple sclerosis data set, configuration files for our tool were smaller by factors of between 3.9 (tranSMART) and 216.1 (i2b2). For the billing data set, configuration files where smaller by factors of between 1.2 (tranSMART) and 1135.0 (i2b2). We note that the sizes were (roughly) equal only for the billing data set and tranSMART, which is because this data set is highly structured and because this type of data is well supported by tranSMART. We conclude that our tool can significantly reduce the efforts required for configuring the loading process.

Table 2. Comparison of input required for data loading.

Data set	Microbiome profiles	Multiple sclerosis	Billing data
LOC[a] input	496	1090	83
LOC staging, i2b2[b]	8772	235,582	94,213
LOC staging, tranSMART	10,976	4272	99
Input files	15	19	11
Staging files, i2b2	2207	1034	31
Staging files, tranSMART	2194	854	18

[a]LOC: lines of configuration.

[b]i2b2: Informatics for Integrating Biology and the Bedside.

Experiment 3: Scalability

In this section, we present the results on the scalability of our tools with respect to increasing volumes of data. The execution times measured in the experiments are provided in Table 3.

The table shows the time needed for staging and loading the data from the 3 evaluation data sets for i2b2 and tranSMART. As can be seen, the execution times scaled roughly linearly with the number of facts loaded into the target systems. Moreover, the relative time needed for data staging was the highest for the multiple sclerosis data set, which is also the data set with the highest complexity, thus requiring the most preprocessing.

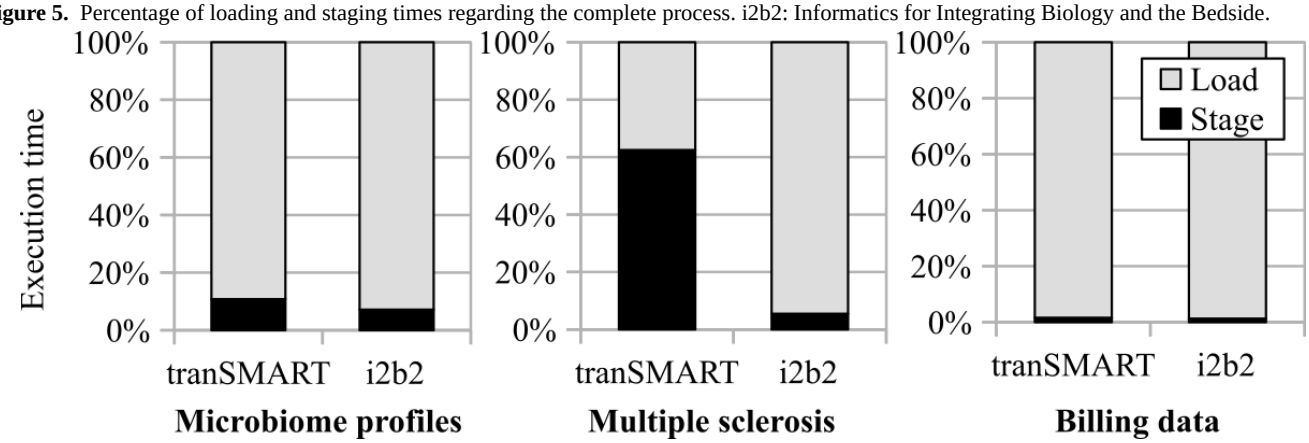
Figure 5 provides an overview of the relationship between the times needed for staging and loading. As can be seen, the (relative) staging times for tranSMART were generally higher than those for i2b2. This can be explained by the fact that more data normalization and restructuring were needed to be performed by the tool to ensure that the data could be loaded into the target system. In addition, more complicated procedures for duplicate detection were needed, as there is little support for the time axis in tranSMART. In summary, we conclude that our approach is scalable and can be used to process large data sets.

Table 3. Execution times of data-loading processes in seconds.

Data set	Microbiome profiles	Multiple sclerosis	Billing data
tranSMART			
Staging time	13	687	91
Loading time	109	413	5687
Total time	122	1100	5778
i2b2[a]			
Staging time	11	804	790
Loading time	144	13,895	61,417
Total time	155	14,699	62,208

[a]i2b2: Informatics for Integrating Biology and the Bedside.

Figure 5. Percentage of loading and staging times regarding the complete process. i2b2: Informatics for Integrating Biology and the Bedside.



Discussion

Principal Findings

We have presented a comprehensive cloud-based platform and a flexible data-loading pipeline to enable the agile provisioning of clinical and translational data warehousing solutions. We have presented an extensive experimental evaluation, dealing with different types of data and targeting platforms with different data analytics capabilities. The results of our analysis show that the presented platform significantly simplifies the management of the supported data warehousing solutions and enables quick loading of data in various representations. This enables the development of such platforms in close cooperation with users based on short feedback cycles. The cloud-based hosting infrastructure and the data-loading pipeline are available as open source software.

The infrastructure and tools presented in this study and the data sets used in our experimental evaluation have been used to support a variety of real-world projects. In particular, the infrastructure is being used to support a large clinical research center [40] that studies shifts in the composition and activity of the microbial ecosystem focusing on clinical endpoints that are associated with well-documented changes in the gut microbiome (inflammation and cancer). For this purpose, a platform is being set up to provide researchers with integrated access to different types of data generated within the consortium. Moreover, our platform is being used within the DIFUTURE (Data Integration for Future Medicine) project to improve data availability and accessibility through an integrated view on health care and research resources, such as biobanks [6]. An important example of one of the use cases of the project is the development of an infrastructure for personalized optimal treatment of multiple sclerosis combined with efforts to better understand the disease in general. Finally, the billing data set has been used in a nationwide cross-site analysis aiming at the reproduction of published comorbidity scores and the descriptive analysis and visualization of the distribution of comorbidity scores as well as the distribution of rare diseases in Germany [41].

Comparison With Prior Work

Analytics Platforms

Currently, our solutions support i2b2 version 1.7.09c and tranSMART version 16.3. In future work, we plan to add support for further warehousing platforms and further versions to support further use cases. An important system of interest is i2b2-tranSMART, which is the result of an initiative to integrate tranSMART with the i2b2 cohort selection services and improved support for managing time series data [42]. In theory, this would obviate the need to support 2 different systems (i2b2 and tranSMART) with a similar technological basis. However, i2b2-tranSMART is still under active development and is not yet suitable for deployment in production environments. It is planned to release this software directly as a Docker container; therefore, we expect little effort to integrate it into the presented environment.

The OMOP CDM and OHDSI toolset also provide an interesting target platform [11]. OHDSI is an international collaborative initiative aimed at making clinical data accessible to analytics efforts, also in distributed settings, to generate actionable insights for improving health care. The OMOP CDM is a CDM for consistently representing health care data from diverse sources by making the relationships between different concepts explicit [11]. The OHDSI project provides a wide range of analytics front-ends, such as ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems) or Atlas, an open source application developed as a part of OHDSI intended to provide a unified interface to patient level data and analytics. Both are aimed at end users and can be deployed over the OMOP CDM. Supporting OMOP/OHDSI within the described cloud-based hosting infrastructure will not be too complex. Implementing an agile loading process, however, will be challenging as the OMOP CDM requires a significant amount of data normalization and encoding with standard terminologies. Finally, cBioPortal would be an important additional system to support as it provides a platform for interactive exploration of multi-dimensional genomics data sets, intending to also support rapid, intuitive, and high-quality access to molecular data and clinical data [43]. A dockerized version for the presented cloud environment has already been implemented, but integrating the software with our data-loading pipeline requires more work.

Cloud-Based Infrastructures

Regarding cloud-based management infrastructures for clinical and translational data warehousing, most studies focus on i2b2 only. The *i2b2 Wizard*, which is part of the IDRT, as well as i2b2 Quickstart aims to simplify installation, setup, and

administration of single i2b2 instances. There are also images available on Docker Hub. However, as neither the source code of these images is publicly available in full nor can their authenticity be verified (eg, using Docker Content Trust [DCT]), we could not use them as a base for further development because of security considerations. For tranSMART, a large number of images are available on Docker Hub. However, they have not been maintained for some time, contain artifacts with unclear provenance, or their documentation leaves out important aspects.

We compared these alternative solutions with our approach with respect to the following criteria:

1. *Supported target platforms* indicates whether a solution can be used for the current major version of i2b2 (ie, 1.7.x) and/or tranSMART (ie, 16.3).
2. *Container-based* denotes whether the solution is encapsulated using container virtualization, which significantly increases the ease and robustness of the installation procedures.
3. *Security by default* covers 3 subcriteria—whether *transport encryption* is part of the default deployment, whether the solution automatically provides strong default passwords and whether these can be changed in an integrated way, that is, without risking to break the application (*password management*), and whether the solution uses or provides means to verify the trustworthiness of the installation package, for example, by using digital signatures or by providing the source code (*trusted runtime environment*).
4. *Unified interface* shows whether the solution helps manage multiple warehouse instances of different types.
5. *Sustainability* covers 2 subcriteria—*full availability of source code* is important for customizing the solution to local requirements and the *last update* of the installation package is an indicator of whether the solution is actively maintained by the provider of the solution or by the community.

The results of the comparison are presented in Tables 4-5.

As can be seen, our infrastructure is the only off-the-shelf solution supporting both i2b2 and tranSMART. Moreover, our software, the IDRT i2b2 Wizard, and i2b2 Quickstart are the only solutions that fulfill requirement 1 (robust installation of a trusted runtime environment), as the other (cloud-based) solutions are not capable of providing a trusted runtime environment due to the reasons explained above. However, i2b2 Wizard and i2b2 Quickstart are not container-based solutions but rather script-based solutions and thus are significantly less flexible than our tool, which is based on container virtualization. Furthermore, our tool is the only solution that fulfills requirement 2 (unified installation and maintenance) because it provides integrated support for both i2b2 and tranSMART through common commands. Finally, our tool is the only solution that fulfills requirement 3 (built-in security) as it is the only solution that provides out-of-the-box support for multiple important security features, such as transport encryption and strong passwords. The IDRT i2b2 Wizard is quite outdated and has not received updates in more than 2 years.

Table 4. Comparison of provisioning infrastructures: Our solution, IDRT[a] and i2b2[b] Quickstart.

Feature	Our solution	IDRT[a] [26]	i2b2[b] Quickstart [27]
Supported target platforms			
i2b2 (current major version)	Yes	No	Yes
tranSMART (current major version)	Yes	No	No
Container based	Yes	No	No
Security by default			
Transport encryption	Yes	No	No
Password management	Yes	No	No
Trusted runtime environment	Yes	Yes	Yes
Unified interface			
Central multi-instance management	Yes	No	No
Sustainability			
Full availability of source code	Yes	Yes	Yes
Last update	March 2020	August 2017	February 2020

[a]IDRT: Integrated Data Repository Toolkit.

[b]i2b2: Informatics for Integrating Biology and the Bedside.

Table 5. Comparison of provisioning infrastructures: i2b2[a] on Dockerhub, tranSMART on Dockerhub, and manual installation.

Feature	i2b2[a] on Dockerhub [28]	tranSMART on Dockerhub	Manual installation
Supported target platforms			
i2b2 (current major version)	Yes	No	Yes
tranSMART (current major version)	No	Yes	Yes
Container based	Yes	Yes	No
Security by default			
Transport encryption	No	No	Yes
Password management	No	No	Yes
Trusted runtime environment	No	No	Yes
Unified interface			
Central multi-instance management	No	No	No
Sustainability			
Full availability of source code	No	No	Yes
Last update	February 2020	October 2019	April 2020

[a]i2b2: Informatics for Integrating Biology and the Bedside.

Data-Loading Tools

In addition to transmart-batch and tMDataLoader, which are both used by our solution, there are further data loaders for tranSMART and i2b2. First, transmart-ETL is the standard loading tool for tranSMART. It is included in the standard software installation of tranSMART and is based on the Pentaho Data Integration platform. Second, ICE is a data loading and curation tool supporting a graphical user interface [36]. Third, transmart-copy is a very lightweight loading tool that copies data provided in a tabular form into the tables of the tranSMART database. tmtk is the solution most similar to our approach. It is a Python-based solution that enables the integration of data via a high-level language and several classes. It is typically used in Jupyter notebooks. Analogous to our solution, it uses transmart-batch as a loading tool. It also supports flexible means for organizing data into entities and attributes through an additional graphical tool called the Arborist. Moreover, for i2b2 only, there are other loading tools available. The most comprehensive is the IDRT Import and Mapping Tool [26]. The tool supports various import formats, such as CSV files; provides access to structured query language databases, such as Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) [44,45]; and provides direct support for CDMs, that are, for example, used for billing purposes. Talend Open Studio is used for all ETL processes.

We compared these tools with our approach with respect to the following criteria:

1. As in the previous section, the criterion *supported target platforms* shows whether a solution can be used for the current major version of i2b2 (ie, 1.7.x) and/or tranSMART (ie, 16.3).
2. The criterion *EAV schema support* indicates whether the tool supports EAV input data with multiple attribute columns (*multi-column*) or with only one attribute column (*basic*).
3. *Automated data cleansing and preprocessing* covers subcriteria indicating whether the tool can handle *different encodings*, *data types*, and *syntaxes* for different data sources or if the tool requires all incoming data to conform to a single, predefined specification, and the subsequent subcriteria show whether the tool can handle *missing or invalid data* and *duplicate data* or whether the ETL process is aborted if it encounters one of these anomalies.
4. The criterion *loading strategy* indicates whether the tool employs other data-loading tools (*meta*) or whether the tool implements its own loading procedures (*direct*).
5. *Configuration paradigm* indicates whether the tool configuration follows a declarative approach or an *imperative* approach.
6. The criterion *sustainability*, as in the previous section, covers 2 subcriteria with the same semantics—*full availability of source code* and the *last update*.

The results of the comparison are provided in Tables 6-7.

Table 6. Comparison of extraction-transformation-loading tools: Our solution, tranSMART-ETL[a], tMData-loader, and transmart-batch.

Feature	Our solution	tranSMART-ETL[a] [33]	tMData-loader [34]	transmart-batch [35]
Supported target platforms				
i2b2[b] (current major version)	Yes	No	No	Yes
tranSMART (current major version)	Yes	Yes	Yes	Yes
EAV[c] schema support	Multi-column	Basic	Basic	Basic
Automated data cleansing and preprocessing				
Different encodings, data types, and syntaxes	Yes	No	No	No
Missing or invalid data	Yes	No	No	No
Duplicate data	Yes	Yes	No	No
Loading strategy	Meta	Direct	Direct	Direct
Configuration paradigm	Declarative	Imperative	Imperative	Imperative
Sustainability				
Source code fully available	Yes	Yes	Yes	Yes
Last update	March 2020	March 2018	December 2017	June 2016

[a]ETL: extraction-transformation-loading.

[b]i2b2: Informatics for Integrating Biology and the Bedside.

[c]EAV: entity-attribute-value.

Table 7. Comparison of extraction-transformation-loading tools: Integrated Curation Environment, Integrated Data Repository Toolkit, transmart-copy, and tmtk[a].

Feature	ICE[b] [36]	IDRT[c] [26]	tranSMART-copy [37]	tmtk[a] [38]
Supported target platforms				
i2b2[d] (current major version)	No	No	No	No
tranSMART (current major version)	Yes	No	Yes	Yes
EAV[e] schema support	Basic	No	No	Basic
Automated data cleansing and preprocessing				
Different encodings, data types, and syntaxes	No	No	No	No
Missing or invalid data	No	No	No	Yes
Duplicate data	No	Yes	No	No
Loading strategy	Meta	Direct	Direct	Meta
Configuration paradigm	Imperative	Imperative	Imperative	Imperative
Sustainability				
Source code fully available	No	Yes	Yes	Yes
Last update	July 2016	August 2017	December 2019	February 2020

[a]tmtk: TranSMART data curation toolkit.

[b]ICE: Integrated Curation Environment.

[c]IDRT: Integrated Data Repository Toolkit.

[d]i2b2: Informatics for Integrating Biology and the Bedside.

[e]EAV: entity-attribute-value.

As can be seen, our solution and transmart-batch are the only tools to support both i2b2 and tranSMART and thus to fulfill requirement 1 (*platform independence*). Requirement 2 (*support for different types of data*) is strongly connected to requirement 3 (*automated data cleansing and preprocessing*). At the structural level, our tool is the only tool to support EAV schema resolution in which multiple columns can be combined (eg, *lab analytes* together with *units of measurement*) and thus is the only one to fulfill requirement 2 (*support for different types of data*). Moreover, our tool is also the only one that is capable of

automatically detecting and handling multiple input data properties, such as encodings, syntaxes, and data types, and thus to ingest heterogeneous data often encountered in the clinical context. Our tool and tranSMART-ETL are both capable of automatically handling duplicate data. In addition to our tool, tmtk and ICE are also meta-loading tools; however, they have fewer data cleansing functionalities. tMDataLoader, ICE, and IDRT are quite outdated and have not received updates in more than 1.5 years.

We conclude that our set of tools is the only solution that supports all requirements outlined in the *Methods* section. Moreover, our solutions are fully open source software, allowing users to maintain their own version if needed, thus decreasing the risks of adoption and improving sustainability.

Limitations and Future Work

In future work, we plan to address the limitations of the current version of the infrastructure. First, the current implementation does not scale to huge data volumes. At the infrastructure level, this would require support for shared databases. On the data-loading layer, support for processing data in the form of smaller blocks or chunks is needed. Extending the data-loading pipeline with this feature will be relatively straightforward. However, the loading tools used as backends need to support incremental loading, which is currently only supported for i2b2 with the transmart-batch backend. In general, the pipeline would benefit significantly from incremental loading capabilities; therefore, we are exploring options to integrate an incremental loading procedure directly into the software.

An additional area of future improvements is authentication and authorization management. For deployments with a large user base, the use of single sign-on concepts, such as OAuth2 [46], will become relevant. As tranSMART uses Spring Security [47], which supports OAuth2, this should be straightforward to accomplish. However, the software stack used by i2b2 does not support OAuth2 natively. Therefore, we plan to evaluate the approach described by Wagholikar et al [48]. Another limitation in terms of information security is that our use of DCT [31] is currently restricted to checking the authenticity and integrity of the base images when building the images. In future versions, we plan to use DCT to sign images as well, which is particularly important when publishing them on the internet.

The current version of the infrastructure focuses on clinical data or selected genomic variants. TranSMART, however, has built-in support for a wide range of high-dimensional data types (see the *Selection of Target Systems* section). In future work, we plan to add support for loading these types of data as well. Although this will require some effort, such data are typically much more structured and represented in standardized formats than the data considered in this study.

Currently, our loading pipeline focuses on automated structural and syntactic harmonization. Automated mapping procedures to standard terminologies are not yet implemented, mainly because in a first step, we have developed the pipeline following our project-specific requirements. Here, all data sets integrated until now have mostly either been (1) collected in a structured form, using standard terminologies as they were captured; (2) mapped to standard terminologies before they were fed into our pipeline; or (3) loaded for use cases that did not require mapping to semantic standards. However, semantic harmonization is a very important process, and the implementation of interfaces to terminology and ontology services directly into our pipeline is part of our development roadmap.

Finally, we also plan to integrate a wide range of privacy-enhancing technologies into the pipeline. In previous work, we have already integrated a flexible method for data anonymization into an earlier version of our software [49]. Currently, we are working on integrating the pipeline with a HL7 FHIR (Health Level Seven Fast Healthcare Interoperability Resources)–based pseudonymization component.

Summary and Conclusions

In this paper, we have presented a flexible infrastructure that supports the agile development and provisioning of translational data analytics platforms to researchers. Our solution helps to bridge the interdisciplinary gap between clinicians and informaticians by enabling the creation of data warehousing solutions in an iterative process involving short feedback cycles following a pay-as-you-go approach [15]. We have achieved this by combining a Docker-based (private) cloud infrastructure for managing warehouse instances with a flexible and easy-to-use loading pipeline based on a declarative configuration paradigm. We have used the platform successfully to support a wide range of projects that used different types of data, which we used in our experiments. The solutions described in this paper are available to the community as open source software [19,20].

Acknowledgments

The authors wish to thank the reviewers for their insightful comments, which helped to significantly improve the earlier versions of this manuscript. The work was, in parts, funded by the German Federal Ministry of Education and Research within the *Medical Informatics Funding Scheme* under reference number 01ZZ1804A (Data Integration for Future Medicine).

Conflicts of Interest

None declared.

References

1. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 2011 Mar;8(3):184-187. [doi: 10.1038/nrclinonc.2010.227] [Medline: 21364692]
2. Schneeweiss S. Learning from big health care data. N Engl J Med 2014 Jul 5;370(23):2161-2163. [doi: 10.1056/NEJMp1401111] [Medline: 24897079]
3. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
4. Tran BX, McIntyre RS, Latkin CA, Phan HT, Vu GT, Nguyen HL, et al. The current research landscape on the artificial intelligence application in the management of depressive disorders: a bibliometric analysis. Int J Environ Res Public Health 2019 Jun 18;16(12):- [FREE Full text] [doi: 10.3390/ijerph16122150] [Medline: 31216619]
5. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010 Dec 10;2(57):57cm29. [doi: 10.1126/scitranslmed.3001456] [Medline: 21068440]
6. Prasser F, Kohlbacher O, Mansmann U, Bauer B, Kuhn KA. Data integration for future medicine (DIFUTURE). Methods Inf Med 2018 Jul;57(S 01):e57-e65 [FREE Full text] [doi: 10.3414/ME17-02-0022] [Medline: 30016812]
7. Kimball R, Caserta J. Surronding the requirements. In: The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Hoboken, New Jersey, USA: John Wiley & Sons; 2011:3-28.
8. Halevy A, Korn F, Noy N, Olston C, Polyzotis N, Roy S, et al. Goods: Organizing Google's Datasets. In: Proceedings of the 2016 International Conference on Management of Data. 2016 Presented at: SIGMOD'16; June 26-July 1, 2016; San Francisco, CA, USA. [doi: 10.1145/2882903.2903730]
9. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
10. Scheufele E, Aronzon D, Coopersmith R, McDuffie MT, Kapoor M, Uhrich CA, et al. tranSMART: an open source knowledge management and high content data analytics platform. AMIA Jt Summits Transl Sci Proc 2014;2014:96-101 [FREE Full text] [Medline: 25717408]
11. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
12. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One 2013;8(3):e55811 [FREE Full text] [doi: 10.1371/journal.pone.0055811] [Medline: 23533569]
13. Killcoyne S, Boyle J. Managing chaos: lessons learned developing software in the life sciences. Comput Sci Eng 2009 Dec;11(6):20-29 [FREE Full text] [doi: 10.1109/MCSE.2009.198] [Medline: 20700479]
14. Kannan V, Basit MA, Youngblood JE, Bryson TD, Toomay SM, Fish JS, et al. Agile co-development for clinical adoption and adaptation of innovative technologies. Health Innov Point Care Conf 2017 Nov;2018:56-59 [FREE Full text] [doi: 10.1109/HIC.2017.8227583] [Medline: 30364762]
15. Franklin M, Halevy A, Maier D. From databases to dataspaces. SIGMOD Rec 2005 Dec;34(4):27-33. [doi: 10.1145/1107499.1107502]
16. Prasser P. Incremental Ontology-Based Integration for Translational Medical Research. Technical University of Munich. 2013. URL: https://mediatum.ub.tum.de/doc/1119200/document.pdf [accessed 2020-05-31]
17. Petrović M, Vučković M, Turajlić N, Babarogić S, Aničić N, Marjanović Z. Automating ETL processes using the domain-specific modeling approach. Inf Syst E-Bus Manage 2016 Jul 9;15(2):425-460. [doi: 10.1007/s10257-016-0325-8]
18. Dingsøyr T, Nerur S, Balijepally V, Moe NB. A decade of agile methodologies: towards explaining agile software development. J Syst Software 2012 Jun;85(6):1213-1221. [doi: 10.1016/j.jss.2012.02.033]
19. Spengler H. Analytics Environment. DIFUTURE. 2020. URL: https://gitlab.com/DIFUTURE/analytics-environment [accessed 2020-05-31]
20. Spengler H, Lang C, Mahapatra T, Gatz I, Prasser F. ETL Pipeline. DIFUTURE. 2020. URL: https://gitlab.com/DIFUTURE/etl-pipeline [accessed 2020-05-31]
21. Canuel V, Rance B, Avillach P, Degoulet P, Burgun A. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. Brief Bioinform 2015 Mar;16(2):280-290 [FREE Full text] [doi: 10.1093/bib/bbu006] [Medline: 24608524]
22. Tatonetti NP, Denny JC, Murphy SN, Fernald GH, Krishnan G, Castro V, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clin Pharmacol Ther 2011 Jul;90(1):133-142 [FREE Full text] [doi: 10.1038/clpt.2011.83] [Medline: 21613990]
23. Jannot A, Zapletal E, Avillach P, Mamzer M, Burgun A, Degoulet P. The Georges Pompidou University hospital clinical data warehouse: a 8-years follow-up experience. Int J Med Inform 2017 Jun;102:21-28. [doi: 10.1016/j.ijmedinf.2017.02.006] [Medline: 28495345]
24. Geerts H, Dacks PA, Devanarayan V, Haas M, Khachaturian ZS, Gordon MF, Brain Health Modeling Initiative (BHMI). Big data to smart data in Alzheimer's disease: the brain health modeling initiative to foster actionable knowledge. Alzheimers Dement 2016 Sep;12(9):1014-1021 [FREE Full text] [doi: 10.1016/j.jalz.2016.04.008] [Medline: 27238630]

25. i2b2 Installation Guide. i2b2 Community Wiki. 2020. URL: https://community.i2b2.org/wiki/display/getstarted/i2b2+Installation+Guide [accessed 2020-05-31]
26. Bauer CR, Ganslandt T, Baum B, Christoph J, Engel I, Löbe M, et al. Integrated data repository toolkit (IDRT). A suite of programs to facilitate health analytics on heterogeneous medical data. Methods Inf Med 2016;55(2):125-135. [doi: 10.3414/ME15-01-0082] [Medline: 26534843]
27. Wagholikar KB, Mendis M, Dessai P, Sanz J, Law S, Gilson M, et al. Automating installation of the integrating biology and the bedside (i2b2) platform. Biomed Inform Insights 2018;10:1178222618777749 [FREE Full text] [doi: 10.1177/1178222618777749] [Medline: 29887730]
28. Wagholikar KB, Dessai P, Sanz J, Mendis ME, Bell DS, Murphy SN. Implementation of informatics for integrating biology and the bedside (i2b2) platform as Docker containers. BMC Med Inform Decis Mak 2018 Jul 16;18(1):66 [FREE Full text] [doi: 10.1186/s12911-018-0646-2] [Medline: 30012140]
29. Internet Engineering Task Force. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. Request for Comments. 2008. URL: https://tools.ietf.org/html/rfc5280#section-4.2.1.6 [accessed 2020-05-31]
30. Souppaya M, Morello J, Scarfone K. Application Container Security Guide. NIST Special Publication 800-190. 2017. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-190.pdf [accessed 2020-05-31]
31. Content Trust in Docker. Docker Documentation. 2020. URL: https://docs.docker.com/engine/security/trust/content_trust/ [accessed 2020-05-31]
32. T'so T. pwgen. GitHub. 2018. URL: https://github.com/tytso/pwgen [accessed 2020-05-31]
33. transmart-ETL. GitHub. 2018. URL: https://github.com/transmart/tranSMART-ETL [accessed 2020-05-31]
34. tMDataLoader. GitHub. 2020. URL: https://github.com/Clarivate-LSPS/tMDataLoader [accessed 2020-05-31]
35. tranSMART Batch. GitHub. 2016. URL: https://github.com/tranSMART-Foundation/transmart-batch [accessed 2020-05-31]
36. transmart-ICE. GitHub. 2016. URL: https://github.com/transmart/transmart-ICE [accessed 2020-05-31]
37. The Hyve. transmart-copy. GitHub. 2019. URL: https://github.com/thehyve/transmart-core/tree/dev/transmart-copy [accessed 2020-05-31]
38. The Hyve. tmtk. GitHub. 2020. URL: https://github.com/thehyve/tmtk/ [accessed 2020-05-31]
39. Lloyd J. Practical Advantages of Declarative Programming. In: Joint Conference on Declarative Programming. 1994 Presented at: GULP-PRODE'94; September 19-22, 1994; Peñiscola, Spain.
40. Haller D. Microbiome Signatures. CRC 1371. 2019. URL: https://www.sfb1371.tum.de/ [accessed 2020-05-31]
41. Kamdje-Wabo G, Gradinger T, Löbe M, Lodahl R, Seuchter SA, Sax U, et al. Towards structured data quality assessment in the german medical informatics initiative: initial approach in the MII demonstrator study. Stud Health Technol Inform 2019 Aug 21;264:1508-1509. [doi: 10.3233/SHTI190508] [Medline: 31438205]
42. I2b2 tranSMART Foundation. tranSMART PMC Roadmap. Google Docs. 2018. URL: http://roadmap-i2b2-transmart-pmc.hms.harvard.edu [accessed 2020-05-31]
43. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013 May 2;6(269):pl1 [FREE Full text] [doi: 10.1126/scisignal.2004088] [Medline: 23550210]
44. Kubick WR, Ruberg S, Helton E. Toward a comprehensive CDISC submission data standard. Drug Inf J 2016 Aug 28;41(3):373-382. [doi: 10.1177/009286150704100311]
45. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC operational data model standard: a methodological review. J Biomed Inform 2016 May;60:352-362 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.016] [Medline: 26944737]
46. The OAuth 2.0 Authorization Framework. IETF Tools. 2012. URL: https://tools.ietf.org/html/rfc6749 [accessed 2020-05-31]
47. Nachimuthu N. Spring Security OAuth. Spring Projects. 2016. URL: https://spring.io/projects/spring-security-oauth [accessed 2020-05-31]
48. Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG, et al. SMART-on-FHIR implemented over i2b2. J Am Med Inform Assoc 2017 Mar 1;24(2):398-402 [FREE Full text] [doi: 10.1093/jamia/ocw079] [Medline: 27274012]
49. Prasser F, Spengler H, Bild R, Eicher J, Kuhn KA. Privacy-enhancing ETL-processes for biomedical data. Int J Med Inform 2019 Jun;126:72-81 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.03.006] [Medline: 31029266]

Abbreviations

ACHILLES: Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems
CDISC: Clinical Data Interchange Standards Consortium
CPU: central processing unit
CSV: comma-separated values
CDM: common data model
DCT: Docker Content Trust
DIFUTURE: Data Integration for Future Medicine

EAV: entity-attribute-value
ER: entity-relationship
ETL: extraction-transformation-loading
HL7 FHIR: Health Level Seven Fast Healthcare Interoperability Resources
i2b2: Informatics for Integrating Biology and the Bedside
ICE: Integrated Curation Environment
IDRT: Integrated Data Repository Toolkit
IT: information technology
ODM: Operational Data Model
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
tmtk: TranSMART data curation toolkit

Edited by C Lovis; submitted 17.09.19; peer-reviewed by E Frontoni, R Ho, E Andrikopoulou, Z He; comments to author 05.12.19; revised version received 16.02.20; accepted 06.05.20; published 21.07.20.

Please cite as:
Spengler H, Lang C, Mahapatra T, Gatz I, Kuhn KA, Prasser F
Enabling Agile Clinical and Translational Data Warehousing: Platform Development and Evaluation
JMIR Med Inform 2020;8(7):e15918
URL: https://medinform.jmir.org/2020/7/e15918
doi:10.2196/15918
PMID:32706673

Original Paper

Exploring the Determinants of Mobile Health Adoption by Hospitals in China: Empirical Study

Boumediene Ramdani[1], PhD; Binheng Duan[2], PhD; Ilhem Berrou[3], PhD

[1]Centre for Entrepreneurship, College of Business & Economics, Qatar University, Doha, Qatar

[2]Creative Assembly, Spire Court, Albion Way, Horsham, United Kingdom

[3]Faculty of Health and Applied Sciences, University of the West of England, Bristol, United Kingdom

Corresponding Author:
Boumediene Ramdani, PhD
Centre for Entrepreneurship
College of Business & Economics
Qatar University
PO Box 2713
Doha
Qatar
Phone: 974 44037762
Email: B.Ramdani@qu.edu.qa

Abstract

Background: Although mobile health (mHealth) has the potential to transform health care by delivering better outcomes at a much lower cost than traditional health care services, little is known about mHealth adoption by hospitals.

Objective: This study aims to explore the determinants of mHealth adoption by hospitals using the technology-organization-environment (TOE) framework.

Methods: We conducted an interviewer-administered survey with 87 managers in Chinese public hospitals and analyzed the data using logistic regression.

Results: The results of our survey indicate that perceived ease of use (β=.692; *P*<.002), system security (β=.473; *P*<.05), top management support (β=1.466; *P*<.002), hospital size (β=1.069; *P*<.004), and external pressure (β=.703; *P*<.005) are significantly related to hospitals' adoption of mHealth. However, information technology infrastructure (β=.574; *P*<.02), system reliability (β=−1.291; *P*<.01), and government policy (β=2.010; *P*<.04) are significant but negatively related to hospitals' adoption of mHealth.

Conclusions: We found that TOE model works in the context of mHealth adoption by hospitals. In addition to technological predictors, organizational and environmental predictors are critical for explaining mHealth adoption by Chinese hospitals.

(JMIR Med Inform 2020;8(7):e14795) doi:10.2196/14795

KEYWORDS

mHealth; mobile phone; adoption; hospitals; TOE; China

Introduction

Background

The aging population and the high prevalence of complex long-term conditions are placing unprecedented pressure on hospital services in China [1,2]. Mobile health (mHealth) not only has the potential to alleviate pressure on hospital services but can also increase accessibility and meet individual patient demands. It has been advocated as a complementary approach to traditional (ie, offline) health care services [3]. With over 1.3 billion mobile subscribers [4], mHealth services in China are considered the largest market in the world, accounting for 12.53 billion yuan or US $1.76 billion (1 yuan = US $0.14), in 2017 [5]. Defined as "the use of mobile devices—such as mobile phones, patient monitoring devices, personal digital assistants (PDAs) and wireless devices—for medical and public health practice" [6], mHealth has the potential to transform health care by delivering better outcomes at a lower cost [7]. For patients, mHealth has the potential to improve the health and well-being of individuals by recognizing behaviors, providing a rapid diagnosis of medical conditions, delivering just-in-time interventions, and continuous monitoring of their health status

[8]. Recent evidence shows that mHealth could improve patient experience [9]. For health care providers, mHealth could reduce demands on clinicians' time by minimizing office visits for the management of common conditions and enabling patient self-management [7].

China is a particularly interesting context for this study. In 2015, 1 in 4 persons aged ≥60 years lived in China, making it the largest population of older citizens in the world [10]. This trend is projected to grow by 71% between 2015 and 2030. Moreover, medical institutions in China are concentrated in cities, making it difficult to deliver health care services to the rural population [11]. Chinese policymakers need to address these challenges and find an effective solution that reaches both the elderly and rural populations. For that, mHealth is part of a national strategy to resolve the "difficulty and expense of seeking a doctor" [12].

Chinese hospitals have begun to take up mHealth to deliver health care services [3,9,12,13]. Most emerging literature focuses on patients' adoption of mHealth services. However, little is known about the hospitals' adoption of mHealth. Therefore, this study aims to address this gap by exploring the determinants of mHealth adoption by hospitals.

A large body of research has explored the determinants of adopting health care technologies. As highlighted in Table 1, previous research explored various health care technologies using several theoretical lenses in different settings. From reviewing the literature, it is still unclear what determines the adoption of mHealth by hospitals.

Table 1. Adoption of health care technologies.

Adoption theory	Adoption of technology[a]	Constructs/factors[b]	Method	Data	Location [reference]
TAM[c] with trust and perceived risks	mHealth[d]	Perceived usefulness, perceived ease of use, perceived risk, performance risk, legal concerns, and trust	SEM[e]	388 patients in large hospitals	China [12]
TAM and UTAUT[f]	Medical dashboard system	Performance expectancy, effort expectancy, social influence, and facilitating conditions	SEM	383 physicians and nurses in a tertiary teaching hospital	Korea [14]
TAM and UTAUT	Mobile electronic medical record	Performance expectancy, effort expectancy, attitude, social influence, facilitating conditions, and behavior intention to use	SEM	449 subjects (65 physicians and 385 nurses) in a large tertiary hospital	Korea [15]
UTAUT 2	mHealth	Performance expectancy, effort expectancy, social influence, facilitating conditions, hedonic motivation, price value, habit, waiting time, and self-concept	Factor analysis and path analysis	A total of 3 surveys with 387, 359, and 375 patients who were offered mHealth as an alternative to traditional hospital services	United States, Canada, and Bangladesh [16]
Task-technology fit and social contagion theory	EHR[g]	Authorization, compatibility, data quality, ease of use, information systems relationship, timeliness, locatability, system reliability, and social contagion	SEM	Survey with 51 university students with working experience in the health care sector and used EHR systems in the past	United States [17]
UTAUT	Home health care robots	Performance expectancy, effort expectancy, social influence, facilitating conditions, trust, privacy concerns, ethical concerns, and legal concerns	SEM and power analysis	108 health care professionals and patients working for home health care agencies	United States [18]
Social capital theory, social cognitive theory and TAM	Telehealth	Perceived ease of use, perceived usefulness, system self-efficacy, social participation, institutional trust, and social trust	SEM	365 patients who used a telehealth system for at least one month	Taiwan [19]
TPB[h]	mHealth service	Perceived value, attitude, perceived behavior control, subjective norm, perceived physical condition, resistance to change, technology anxiety, and self-actualization need	SEM	424 middle-aged and older people accessing community service centers	China [20]
UTAUT	HIT[i]	Performance expectancy, effort expectancy, social influence, facilitating conditions, and provincial areas	SEM	400 health care professionals working in hospital	Thailand [21]
No specific theory	mHealth usage intention, assimilation, and channel preferences	Individual difference, health care availability and health care utilization, and socioeconomic status and demographics	Hierarchical ordinary least squares	1132 consumers	United States [22]
Decomposed TPB and value-attitude-behavior	MEDLINE system	Perceived usefulness, perceived ease of use, attitude, interpersonal influence, subjective norm, personal innovativeness in IT[j], self-efficacy, facilitating conditions, perceived behavioral control, and usage intention	SEM	224 physicians in primary care centers and hospitals	Taiwan [23]
TAM and TPB	Mobile health care	Attitude, perceived behavioral control, subjective norm, perceived usefulness, perceived ease of use, personal innovativeness, and perceived service availability	SEM	140 health care professionals working in hospitals	Taiwan [24]
UTAUT	HIT	Performance expectancy, effort expectancy, social influence, voluntariness, facilitating conditions, experience, and IT knowledge	SEM	Information management officers or head officers from 1323 community health centers	Thailand [25]
TAM and innovation diffusion theory	Electronic logistics information system	Compatibility, perceived usefulness, perceived ease of use, trust, perceived financial cost, and behavioral intention	SEM	Nurses working in 10 hospitals who used electronic logistics information system	Taiwan [26]

Adoption theory	Adoption of technology[a]	Constructs/factors[b]	Method	Data	Location [reference]
TAM	Telemedicine	Perceived usefulness and perceived ease of use	SEM	408 physicians working in tertiary hospitals	Hong Kong [27]

[a]Dependent variable.

[b]Independent variables.

[c]TAM: technology acceptance model.

[d]mHealth: mobile health.

[e]SEM: structural equation modelling.

[f]UTAUT: unified theory of acceptance and use of technology.

[g]EHR: electronic health record.

[h]TPB: theory of planned behavior.

[i]HIT: health information technology.

[j]IT: information technology.

Research Model and Hypotheses

The technology-organization-environment (TOE) framework, first introduced by Tornatzky and Fleischer [28], has been used as a guiding theoretical basis for the adoption of many technologies. This framework integrates the 3 vital contexts—technology, organization, and environment—to provide a holistic understanding of adoption of technology from an organizational perspective. The generic nature of the framework allows researchers to explore the determinants that are relevant to their specific context. Moreover, it has the empirical support of several studies exploring the adoption of technology in different types of organizations and different sectors of the economy (Table 2). A number of technologies have been examined from open systems and electronic business (e-business) to enterprise systems, radio-frequency identification (RFID), and cloud computing. Most of these studies focused on the adoption of technology from an organizational perspective, including government agencies [29], large firms [30-32], and small-to-medium–sized enterprises (SMEs) [33-37]. Other studies focused on particular sectors, including hotels [38], retailing [39], manufacturing [38,40], and the services sector [40]. Surprisingly, only two studies addressed the adoption of technology from a hospital perspective [41,42]. This study aims to extend this body of research by using the TOE model to explore the determinants of mHealth adoption by hospitals.

From reviewing the literature, we developed a TOE framework to explore mHealth adoption by hospitals in China. Our framework (Figure 1) suggests that hospital adoption of mHealth is influenced by TOE determinants: technology—perceived usefulness, perceived ease of use, system compatibility, system security, and information technology (IT) infrastructure; organization—top management support (TMS), organizational readiness, and size; and environment—government policy and regulations, and external pressure. Each of the 3 contexts is discussed below to develop our hypotheses.

Table 2. Technology-organization-environment determinants of adoption of technology.

Adoption of technology[a] [reference]	Determinants[b]									
	Technology					Organization			Environment	
	Perceived usefulness	Perceived ease of use	Compatibility	System security	IT[c] infrastructure	Top management support	Organizational readiness	Size	Government policy	External pressure
Cloud computing [29]	N/A[d]	N/A	N/A	N/A	X[e]	X	N/A	N/A	X	X
e-SCM[f] [30]	X	N/A	N/A	N/A	N/A	X	X	N/A	N/A	X
Open systems [32]	N/A	N/A	N/A	N/A	X	N/A	N/A	N/A	N/A	X
Electronic data interchange [33]	X	N/A	N/A	N/A	X	N/A	N/A	N/A	X	X
Internet [34]	X	N/A	N/A	N/A	X	N/A	N/A	N/A	N/A	N/A
Enterprise systems [35]	X	N/A	N/A	N/A	N/A	X	X	X	N/A	N/A
Enterprise applications [36]	X	X	X	N/A	N/A	X	X	X	N/A	X
Enterprise resource planning [37]	X	N/A	X	X	X	N/A	N/A	X	X	X
Mobile reservation systems [38]	X	X	X	N/A	X	X	N/A	X	N/A	X
e-Business[g] [39]	N/A	N/A	N/A	N/A	X	N/A	N/A	X	X	X
Cloud computing [40]	X	X	N/A	N/A	N/A	X	N/A	X	N/A	N/A
Health information exchange [41]	N/A	N/A	N/A	N/A	X	N/A	N/A	N/A	N/A	X
RFID[h] [42]	X	N/A	X	X	N/A	X	X	N/A	X	X
Cloud computing [43]	N/A	X	N/A	N/A	X	N/A	N/A	N/A	N/A	X
RFID [44]	X	N/A	N/A	N/A	N/A	X	N/A	X	N/A	N/A
RFID [45]	X	X	X	N/A	X	X	N/A	X	N/A	X
e-Business [46]	X	N/A	X	X	X	N/A	N/A	X	N/A	X
e-Business [47]	N/A	N/A	N/A	N/A	X	X	N/A	X	X	X
e-Commerce[i] [48]	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	X	X
e-Commerce [49]	N/A	N/A	N/A	N/A	X	N/A	N/A	N/A	X	X

[a]Dependent variable.

[b]Independent variables.

[c]IT: information technology.

[d]N/A: not applicable.

[e]X: significant determinant.

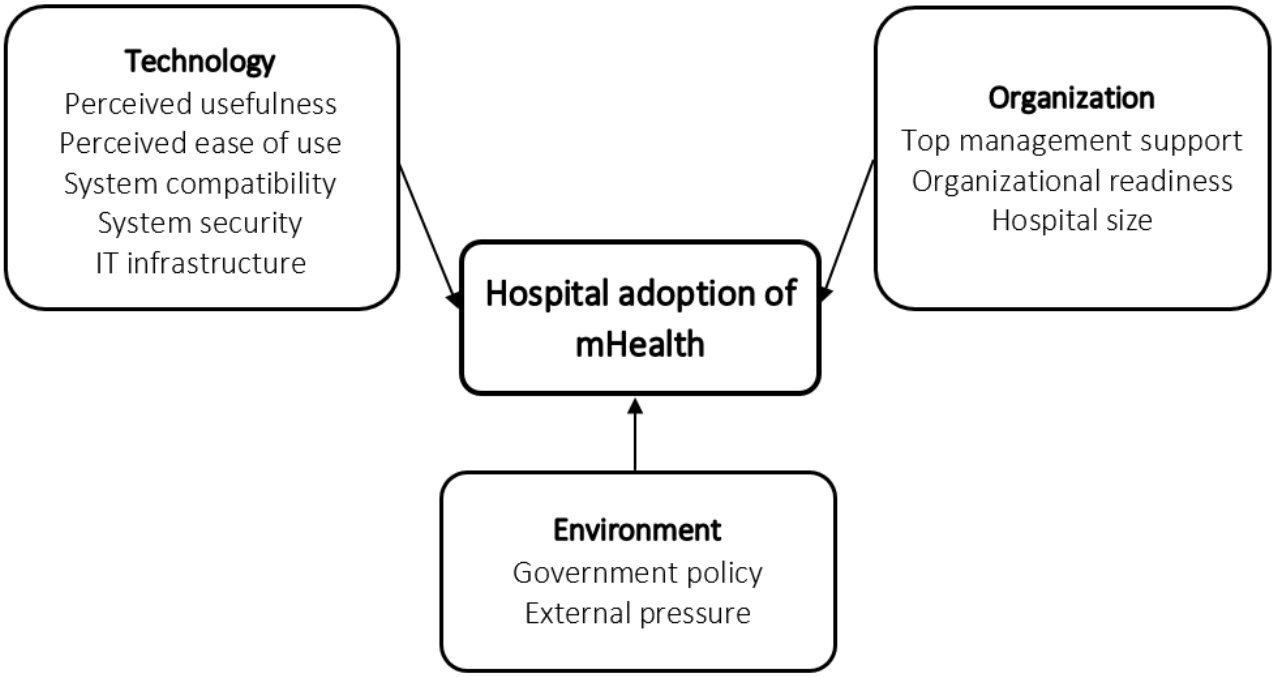
[f]e-SCM: electronic supply chain management.

[g]e-Business: electronic business.

[h]RFID: radio-frequency identification.

[i]e-Commerce: electronic commerce.

Figure 1. Technological organizational and environmental determinants of mHealth adoption by hospitals.



Technology

A number of technological determinants have been shown to affect organizational adoption of new technologies, including perceived usefulness, perceived ease of use, system compatibility, system security, and IT infrastructure. New technology is more likely to be adopted when an organization perceives it to offer more benefits than existing systems. Several new technologies have proven their advantage over existing practices, including e-business over physical stores [31,46], RFID over manual data entry [44,45], cloud computing over client-server computing [40], and mobile reservation over the telephone or web-based reservations [38]. In the health care context, the perceived benefits of RFID in keeping track of hospital patients were found to be a significant determinant of adoption [42]. In their study of customer relationship management (CRM), Hung et al [23] also found a relative advantage to be positively associated with CRM adoption by hospitals. On this basis, we suggested the following:

> *Hypothesis 1: Perceived usefulness will positively influence hospitals' adoption of mHealth.*

Perceived ease of use has been found to be a significant factor in the adoption of technology [36,38,40,43,45]. The widespread use of smartphones and health monitoring devices has made it easier for consumers to handle such devices remotely [16]. For hospitals to have a more active engagement in health care delivery, we expect the perceived ease of use of mobile technologies to be positively associated with mHealth adoption. Thus, we proposed the following:

> *Hypothesis 2: Perceived ease of use will positively influence hospitals' adoption of mHealth.*

System compatibility has been found as a significant determinant for the adoption of technology [36-38,45,46]. In the health care context, compatibility with a non-RFID–based patient tracking system was found to be a crucial determinant for RFID application integration [42]. Therefore, we suggested the following hypothesis:

> *Hypothesis 3: System compatibility with existing systems will positively influence hospitals' adoption of mHealth.*

Unauthorized data access is a concern for organizations and their clients and has the potential to jeopardize information security and privacy [50]. Similar to e-business, mobile technologies are integrated into transactions that involve fund transfer and the exchange of organizational data [46]. Although a few studies [37,46] found system security to be positively associated with the adoption of technology, it is of particular significance for health care providers. Security and privacy protection were found to be major determinants of RFID adoption by hospitals [42].A security breach of patient information not only puts patients at risk but can also cause a lasting damage to a hospital's reputation. Thus, we proposed the following:

> *Hypothesis 4: System security will positively influence hospitals' adoption of mHealth.*

Finally, IT infrastructure has been found to be one of the most significant determinants of the adoption of technology. This factor was found to be significant for most types of technologies, including open systems [32], e-business [31,46], RFID [45], enterprise resource planning [37], mobile reservation systems [38], and cloud computing [29]. In a study of health information exchange (HIE), hospitals that do not have the necessary technological infrastructure were found to be less likely to adopt new systems [41]. Therefore, we expect IT infrastructure to be a significant determinant of mHealth adoption by hospitals, and we proposed the following hypothesis:

> *Hypothesis 5: IT infrastructure will positively influence hospitals' adoption of mHealth.*

Organization

TMS, organizational readiness, and organizational size have been found to influence the organizational adoption of new technologies. TMS has been found to be one of the most significant determinants of the adoption of technology [51] as managers can overcome barriers to adoption and resistance to change [30]. It has been found to be an influential factor in the adoption of e-business [46], enterprise systems [35], RFID [44,45], and mobile reservation systems [38]. In hospitals, TMS has been shown to be a significant determinant of RFID adoption for patient tracking [42]. Thus, we suggested the following:

> *Hypothesis 6: TMS will positively influence hospitals' adoption of mHealth.*

Organizational readiness, the degree to which an organization has the knowledge and resources that can remove barriers to system adoption [52], has been shown to be positively related to the adoption of new technology [30]. It has been found to influence the adoption of enterprise systems significantly [35,36]. In the health care context, organizational readiness has been shown to be of critical importance in RFID adoption [42]. Therefore, we proposed the following:

> *Hypothesis 7: Organizational readiness positively influences hospitals' adoption of mHealth.*

Organizational size is another important determinant of adoption of technology [51]. It was established that larger firms are more likely to adopt enterprise systems [35-37]. Other studies also confirm the significance of organizational size [38-40,45-47]. Thus, we suggested the following:

> *Hypothesis 8: Hospital size will positively influence hospitals' adoption of mHealth.*

Environment

Government policy and external pressure have been shown to influence the organizational adoption of new technologies. Existing laws and regulations can critically impact the adoption of new technologies [29,37]. In the health care context, compliance with legislation and standards has been found to be critical in the adoption of RFID patient tracking [42]. Government regulation has been advocated to play a more important role in the Chinese context compared with other developed economies [49]. As hospitals in China are state-owned, government policy can encourage or discourage hospitals from adopting mHealth. On this basis, we proposed the following:

> *Hypothesis 9: Government policy will positively influence hospitals' adoption of mHealth.*

External pressure has been found to be one of the most significant determinants of organizational adoption of new technologies [51]. Health care organizations are under constant pressure to adopt and implement new technologies to become more efficient [53]. To coordinate care and steer patients away from emergency departments, hospitals are under pressure to adopt HIE [41]. Thus, we suggested the following:

> *Hypothesis 10: External pressure will positively influence hospitals' adoption of mHealth.*

Methods

Data Collection

An interviewer-administered survey was conducted to test the research model empirically. The questionnaire was piloted and refined through rigorous pretesting. In this phase, 8 public hospital managers were invited to participate in this study and comment on instrument clarity and question wording. As a result, construct measures were revised. In addition, system reliability [54] was emphasized by hospital directors as a missing variable that should be included in the research instrument. Here, we revised the research instrument to include system reliability as a further measure of the technology context. A company would choose not to adopt cloud computing because of the increased business risk associated with the uncertainty of service availability and reliability, especially if there are unexpected downtimes and disruptions [55]. People often would not prefer to use new technology because of concerns about the reliability and stability of the system [55]. System reliability is key when providing uninterrupted services as shown in a study by Pagani [56]. Thus, we expect system reliability to influence hospitals' adoption of mHealth significantly.

A convenient snowballing approach was used in this study. Due to the difficulty in obtaining data from public hospitals in China, the authors focused on collecting data from 2 regions, namely Shanghai and Gansu. A total of 91 questionnaires were obtained, but 4 questionnaires were discarded because of missing data. As a result, 87 responses were included in the analysis. Respondents, from hospitals that are not willing to adopt mHealth, were classified as nonadopters, whereas respondents from hospitals that are willing to adopt mHealth in the next 3 years were classified as adopters. Table 3 shows the characteristics of the responding hospitals in terms of location, hospital level, hospital beds, and respondents' positions. Hospitals in China are classified into 3 major tiers, with 3 subtiers within each major tier [29,57]. Level 3 hospitals provide specialized medical services in several departments, with level 3A hospitals being the most advanced. These hospitals have a minimum capacity of 500 beds. Level 2 hospitals are regional hospitals with 100 to 499 beds providing cross-community medical services. These are smaller in size and less advanced than those in level 3 hospitals. Level 1 hospitals provide basic health care facilities with a capacity of 20 to 99 beds.

The questionnaire was translated into Chinese official language (standard Mandarin) following the conventional forward-then-back-translation approach. This has taken into account local culture and dialect considerations when establishing conceptual equivalence between English and Chinese versions of the instrument [58].

Table 3. Hospitals' and respondents' characteristics.

Demographics	Adopters (n=50), n (%)	Nonadopters (n=37), n (%)
Location		
Shanghai	37 (74)	9 (24)
Gansu	13 (26)	28 (76)
Public hospital level		
<Level 3	35 (70)	7 (19)
Level 3A	29 (58)	5 (14)
Level 3B	6 (12)	2 (5)
>Level 3	15 (30)	30 (81)
Level 2	12 (24)	24 (65)
Level 1	3 (6)	6 (16)
Hospitals beds		
500+	35 (70)	7 (19)
100-499	12 (24)	24 (65)
20-99	3 (6)	6 (16)
Respondents' job titles		
Directors of laboratory services	37 (74)	10 (27)
Directors of IT[a] department	6 (12)	9 (24)
Directors of other departments	7 (14)	18 (49)

[a]IT: information technology.

Measures

Measurement items were developed based on a comprehensive review of the literature and modified to suit the mHealth context in China. All measurement items are listed in Multimedia Appendix 1. The items for perceived usefulness, perceived ease of use, and system compatibility were adopted from Moore and Benbasat [59]. The measure for system security was developed from Kim et al [60], the measure for IT infrastructure was developed from Bhattacherjee and Hikmet [61], and the measure for system reliability was adapted from Goodhue and Thompson [54]. The items for TMS were adapted from Yap et al [62], and the items for organizational readiness were adapted from Grandon and Pearson [63]. The items for government policy were adapted from Chau and Tam [32], and the items for external pressure were adopted from Premkumar and Roberts [64]. A 5-point Likert scale ranging from *strongly disagree* to *strongly agree* was used for all measurement items with the exception of hospital size, which was classified into 3 major tiers [29,57].

Results

Validity and Reliability

The validity of construct measures was assessed using principal component analysis with orthogonal factor rotation. All factor loadings were above 0.5 [65]. Reliability was assessed using Cronbach α. All α coefficients exceeded .7 [65]. As shown in Table 4, factor analysis and α coefficient results indicate adequate validity and reliability of the measures.

The correlation matrix was examined for multicollinearity problems. Table 5 shows that none of the squared correlation coefficients are above the 0.9 level [65]. Table 6 shows that the variance inflation factor values are not greater than the cutoff value of 10 [65], indicating that multicollinearity is not a problem for this study.

Table 4. Factor analysis and reliability assessment.

Constructs and items	PU[a]	PE[b]	SC[c]	SS[d]	ITI[e]	SR[f]	TMS[g]	OR[h]	GP[i]	EP[j]
PU1	.603	__[k]	—	—	—	—	—	—	—	—
PU2	.825	—	—	—	—	—	—	—	—	—
PU3	.727	—	—	—	—	—	—	—	—	—
PU4	.878	—	—	—	—	—	—	—	—	—
PU5	.584	—	—	—	—	—	—	—	—	—
PU6	.821	—	—	—	—	—	—	—	—	—
PE1	—	.621	—	—	—	—	—	—	—	—
PE2	—	.820	—	—	—	—	—	—	—	—
PE3	—	.773	—	—	—	—	—	—	—	—
PE4	—	.631	—	—	—	—	—	—	—	—
PE5	—	.739	—	—	—	—	—	—	—	—
PE6	—	.707	—	—	—	—	—	—	—	—
SC1	—	—	.694	—	—	—	—	—	—	—
SC2	—	—	.836	—	—	—	—	—	—	—
SC3	—	—	.663	—	—	—	—	—	—	—
SS1	—	—	—	.930	—	—	—	—	—	—
SS2	—	—	—	.896	—	—	—	—	—	—
SS3	—	—	—	.832	—	—	—	—	—	—
ITI1	—	—	—	—	.688	—	—	—	—	—
ITI2	—	—	—	—	.859	—	—	—	—	—
ITI3	—	—	—	—	.721	—	—	—	—	—
SR1	—	—	—	—	—	.821	—	—	—	—
SR2	—	—	—	—	—	.956	—	—	—	—
SR3	—	—	—	—	—	.772	—	—	—	—
TMS1	—	—	—	—	—	—	.719	—	—	—
TMS2	—	—	—	—	—	—	.819	—	—	—
TMS3	—	—	—	—	—	—	.897	—	—	—
OR1	—	—	—	—	—	—	—	.877	—	—
OR2	—	—	—	—	—	—	—	.859	—	—
OR3	—	—	—	—	—	—	—	.628	—	—
GP1	—	—	—	—	—	—	—	—	.873	—
GP2	—	—	—	—	—	—	—	—	.880	—
EP1	—	—	—	—	—	—	—	—	—	.702
EP2	—	—	—	—	—	—	—	—	—	.823
EP3	—	—	—	—	—	—	—	—	—	.895
Eigenvalue	2.748	4.330	4.039	2.653	2.367	2.686	2.250	2.152	1.711	1.355
Variance	7.634	12.028	11.221	7.369	6.576	7.460	6.251	5.978	4.752	3.764
Cronbach α	.778	.778	.786	.797	.794	.778	.764	.778	.790	.777

[a]PU: perceived usefulness.

[b]PE: perceived ease of use.

[c]SC: system compatibility.

[d]SS: system security.

[e]ITI: information technology infrastructure.
[f]SR: system reliability.
[g]TMS: top management support.
[h]OR: organizational readiness.
[i]GP: government policy.
[j]EP: external pressure.
[k]Constructs and items.

Table 5. Correlations between independent variables.

Constructs	PU[a]	PE[b]	SC[c]	SS[d]	ITI[e]	SR[f]	TMS[g]	OR[h]	HS[i]	GP[j]	EP[k]
PU	1.000	N/A[l]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
PE	−0.277	1.000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SC	−0.247	0.394	1.000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SS	0.005	0.512	0.278	1.000	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ITI	0.054	−0.440	−0.366	−0.368	1.000	N/A	N/A	N/A	N/A	N/A	N/A
SR	−0.054	−0.734	−0.481	−0.497	0.507	1.000	N/A	N/A	N/A	N/A	N/A
TMS	−0.235	0.659	0.489	0.454	−0.594	−0.694	1.000	N/A	N/A	N/A	N/A
OR	.048	0.456	0.440	0.476	−0.496	−0.612	0.516	1.000	N/A	N/A	N/A
HS	−0.122	0.531	0.384	0.471	−0.202	−0.554	0.567	0.269	1.000	N/A	N/A
GP	0.250	−0.786	−0.579	−0.575	0.470	0.708	−0.824	−0.634	−0.562	1.000	N/A
EP	−0.098	0.469	0.332	0.399	−0.207	−0.496	0.512	0.187	0.575	−0.643	1.000

[a]PU: perceived usefulness.
[b]PE: perceived ease of use.
[c]SC: system compatibility.
[d]SS: system security.
[e]ITI: information technology infrastructure.
[f]SR: system reliability.
[g]TMS: top management support.
[h]OR: organizational readiness.
[i]HS: hospital size.
[j]GP: government policy.
[k]EP: external pressure.
[l]N/A: not applicable.

Table 6. Collinearity statistics.

Constructs	Tolerance	Variance inflation factor
Perceived usefulness	0.638	1.567
Perceived ease of use	0.528	1.896
System compatibility	0.735	1.360
System security	0.841	1.188
Information technology infrastructure	0.742	1.347
System reliability	0.567	1.764
Top management support	0.475	2.106
Organizational readiness	0.564	1.773
Hospital size	0.742	1.347
Government policy	0.564	1.773
External pressure	0.543	1.840

Model Testing

Logistic regression was used as the dependent variable was dichotomous (nonadopters vs adopters). This technique has been utilized in previous studies on the organizational adoption of technologies such as mobile reservation systems [38], electronic supply chain management [30], and enterprise systems [35].

Table 7 shows the results of the logistic regression analysis. The chi-square test was significant (omnibus X^2_{11}=70.4; *P*<.001), and 2 pseudo R^2 values (Cox and Snell R^2=0.55; Nagelkerke R^2=0.74) were satisfactory. Moreover, the research model correctly predicted 81% (30/37) of the nonadopters and 88% (44/50) of the adopters with an overall predictive accuracy of 85% (Table 8). Overall, the research model exhibits an acceptable fit with the data.

Table 7. Results of the logistic regression.

Constructs[a]	β coefficient	Wald statistic	*P* value
Perceived usefulness	−.096	0.424	.51
Perceived ease of use	.692[b]	9.406	.002
System compatibility	.561	3.083	.07
System security	.473[c]	3.828	.05
Information technology infrastructure	−.574[c]	4.784	.02
System reliability	−1.291[c]	6.123	.01
Top management support	1.466[b]	9.614	.002
Organizational readiness	.605	2.170	.14
Hospital size	1.069[b]	8.345	.004
Government policy	−2.010[c]	6.516	.04
External pressure	.703[b]	3.972	.005

[a]Goodness-of-fit: omnibus X^2_{11}=70.4; *P*<.001; −2 log likelihood value=118.658; Cox and Snell R^2=0.55; Nagelkerke R^2=0.74.

[b]*P*<.01.

[c]*P*<.05.

Table 8. Discriminating power.

Observed	Predicted		Percentage correct
	Nonadopters, n (%)	Adopters, n (%)	
Nonadopters	30 (81)	7 (19)	81
Adopters	6 (12)	44 (88)	88
Overall	N/A[a]	N/A	85

[a]N/A: not applicable.

As shown in Table 7, perceived ease of use (β=.692; *P*<.002), system security (β=.473; *P*<.05), TMS (β=1.466; *P*<.002), hospital size (β=1.069; *P*<.004), and external pressure (β=.703; *P*<.005) were significantly related to hospitals' adoption of mHealth. IT infrastructure (β=−.574; *P*<.02), system reliability (β=−1.291; *P*<.01), and government policy (β=−2.010; *P*<.04) were significant but negatively related to hospitals' adoption of mHealth. Thus, hypotheses 2, 4, 6, 8, and 10 are supported. However, perceived usefulness, system compatibility, and organizational readiness did not exhibit a significant relationship with hospitals' adoption of mHealth. Thus, hypotheses 1, 3, and 7 are not supported. These findings are summarized in Table 9.

Table 9. Summary of hypotheses support.

Predictors	Hospital mobile health adoption
Perceived usefulness	Reject
Perceived ease of use	Accept
System compatibility	Reject
System security	Accept
Information technology infrastructure	Reject (significant but negative)
System reliability	Reject (significant but negative)
Top management support	Accept
Organizational readiness	Reject
Hospital size	Accept
Government policy	Reject (significant but negative)
External pressure	Accept

Discussion

Principal Findings

This study explores the determinants of hospitals' adoption of mHealth using the TOE framework. The technological determinants of mHealth adoption by hospitals were examined. Although perceived ease of use and system security are facilitators, IT infrastructure and system reliability are inhibitors of mHealth adoption by hospitals. Perceived ease of use has been found to be a significant determinant of mHealth adoption not only among diabetic patients [66] but also among health care professionals [67]. In addition, security and privacy protection has been found to influence hospital adoption of mHealth [66] and RFID patient tracking [42]. Lack of security was found to be a barrier to telemedicine adoption by physicians [68]. Unexpectedly, IT infrastructure and system reliability were significant but negatively related to hospitals' adoption of mHealth. These results differ from those obtained by Vest [41], who noted that hospitals with low levels of IT infrastructure readiness have lower odds of HIE adoption. They also differ from the results obtained by Shareef et al [66], who found that perceived reliability is positively associated with mHealth adoption. A possible explanation for the negative relationships is the lack of a comprehensive strategy to invest, implement, and use mHealth. Evidence indicates that even among hospitals with established strategies to adopt mHealth solutions, only a few attempt to integrate and align these mHealth solutions with their existing IT systems [69].

Surprisingly, perceived usefulness does not exhibit a positive effect on hospitals' adoption of mHealth. The reason for this insignificant result could be because of the lack of awareness of the benefits of adopting mHealth solutions. This finding is similar to that of Wang et al [45] study of RFID adoption by manufacturing firms and the study [38] of adoption of mobile reservation systems by hotels. In addition, system compatibility does not exhibit a positive effect on hospitals' adoption of mHealth. Here, hospital managers seem to underestimate the significance of system compatibility and the extent to which mHealth is perceived to be consistent with their needs, values, and experiences. A possible explanation for this insignificant result is that new mHealth technologies can be easily integrated with existing systems. This finding is similar to that of both study of enterprise systems adoption by SMEs by Ramdani et al [35] and the study of cloud computing adoption in manufacturing and services firms by Oliveira et al [40].

Organizational determinants of mHealth adoption by hospitals were investigated. As expected, TMS and hospital size are facilitators of mHealth adoption by hospitals. These findings are consistent with Cao et al [42], who found management support to be key to the success of RFID application in hospitals, and Hung et al [23], who found hospital size to be a critical factor in CRM adoption by hospitals. Surprisingly, organizational readiness does not exhibit a significant effect on hospitals' adoption of mHealth. Although lack of financing has been found to be a barrier to the adoption of 3 health information technologies, including electronic health record functionalities, electronic-prescriptions, and telemedicine [68], the reason behind the insignificance of organizational readiness is that mHealth technologies do not require a substantial upfront investment. The cost associated with mHealth tends to be much lower than that of traditional medical services [12].

The environmental determinants of mHealth adoption by hospitals were investigated. Although government policy is an inhibitor, external pressure is a facilitator of mHealth adoption by hospitals. Government policy is significant but negatively associated with hospitals' adoption of mHealth. The lack of an enabling health care policy has been suggested as a barrier to mHealth adoption [70]. Furthermore, the current highly regulated environment in hospitals in China could hinder the adoption of new technologies. As expected, external pressure is positively associated with hospital adoption of mHealth. This result is supported by Cao et al [42] study of RFID adoption in the health care sector, where external pressure was found to be one of the key dimensions of the environmental context. Moreover, competition between health care organizations has been found to be a key determinant of HIE adoption [41].

Study Implications and Limitations

The results of this study provide practical implications for mHealth suppliers and policymakers. First, both perceived ease

of use and system security in the technological context have a significant effect on hospitals' adoption of mHealth. To facilitate hospital adoption, mHealth developers and suppliers need to ensure that the adoption process is relatively simple, and the system is highly secure. Second, both TMS and hospital size in the organizational context have a significant positive effect on hospitals' adoption of mHealth. To get hospitals to adopt mHealth, suppliers need to direct their advertising and promotions toward senior executives who make the final decision to adopt. In addition, mHealth suppliers may need to target larger hospitals because they are likely to invest in such systems. Finally, government policy and external pressure in the environmental context have a significant effect on hospitals' adoption of mHealth. Although external pressure facilitates the adoption of mHealth, existing government policies must be revised to encourage the adoption of new technologies in the health care sector.

Several potential limitations must be considered when interpreting the results of this study. First, this study focuses only on hospitals' adoption of mHealth. The impact of mHealth on hospital performance should be examined to gain a holistic understanding. Second, a set of technological, organizational, and environmental predictors were examined. Future studies may examine whether other predictors may influence hospitals' adoption of mHealth. Third, the collected data are cross-sectional and posited causal relationships could only be inferred. Future studies could collect longitudinal data to determine causal links more explicitly. Fourth, the statistical technique employed (ie, logistic regression) only analyses the relationships between hospitals' adoption of mHealth and their predictors and does not analyze the relationships between the predictors. Future studies could use other statistical techniques to examine the relationships between the predictors and elaborate on the findings of this study. Another important limitation lies in the sample size and type of hospitals studied. Although our sample size was adequate for this study, the findings might vary with larger samples. In addition, because private hospitals are developing rapidly in China, it will be worth examining how our findings compare with private hospitals' adoption of mHealth. Finally, our data are largely dominated by hospitals from 2 regions in China: Shanghai and Gansu. We have overlooked the potential regional differences in our study. Thus, the research model should be tested further using larger samples from other regions or even from other counties to make cross-region or cross-country comparisons.

Conclusions

This study contributes to the literature on organizational mHealth adoption by examining the determinants of mHealth adoption by hospitals. The results indicate that significant predictors of hospitals' adoption of mHealth include perceived ease of use, system security, IT infrastructure, system reliability, TMS, hospital size, external pressure, and government policy. However, perceived usefulness and system compatibility in the technological context and organizational readiness in the organizational context are not significant predictors.

The contributions of this study to research on organizational mHealth adoption are 3-fold. First, previous studies focused on the adoption of mHealth at an individual level, including health care professionals and patients. This study adds important insights to the literature by focusing on the organizational (ie, hospital) adoption of mHealth. Second, limited knowledge exists on the adoption of technology in Chinese health care organizations. This study contributes to the literature by highlighting the context-specific determinants of mHealth adoption. Third, studies of adoption of technology in health care organizations mainly use versions of the unified theory of acceptance and use of technology framework. This study uses the TOE framework to contribute to the adoption of technology in the health care literature by identifying the predictors that influence hospitals to adopt mHealth.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire: Measurement of constructs.
[DOCX File , 14 KB - medinform_v8i7e14795_app1.docx]

References

1. Zhang X, Lai K, Guo X. Promoting China's mhealth market: a policy perspective. Health Policy Technol 2017 Dec;6(4):383-388. [doi: 10.1016/j.hlpt.2017.11.002]
2. Sun J, Guo Y, Wang X, Zeng Q. mHealth for aging China: opportunities and challenges. Aging Dis 2016 Jan;7(1):53-67 [FREE Full text] [doi: 10.14336/AD.2015.1011] [Medline: 26816664]
3. Meng F, Guo X, Peng Z, Lai K, Zhao X. Investigating the adoption of mobile health services by elderly users: trust transfer model and survey study. JMIR Mhealth Uhealth 2019 Jan 8;7(1):e12269 [FREE Full text] [doi: 10.2196/12269] [Medline: 30622092]
4. International Telecommunications Union. 2016. ITU Estimates That at the End of 2019, 53.6 Per Cent of the Global Population, or 4.1 Billion People, Are Using the Internet URL: http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx [accessed 2020-05-11]
5. GSMA. 2018. Review of China Mobile Health Market and Outlook for Future URL: https://www.gsma.com/iot/wp-content/uploads/2013/03/1.

-Mr-%E5%BC%A0%E6%AF%85-ZhangYi-ii-Research-Review-of-China-Mobile-Health-Market-and-Outlook-for-Future.pdf [accessed 2019-05-23]
6. World Health Organization. Global Diffusion of eHealth: Making Universal Health Coverage Achievable. Report of the Third Global Survey on eHealth. Geneva, Switzerland: World Health Organization; 2016.
7. Steinhubl SR, Muse ED, Topol EJ. Can mobile health technologies transform health care? J Am Med Assoc 2013 Dec 11;310(22):2395-2396. [doi: 10.1001/jama.2013.281078] [Medline: 24158428]
8. Kumar S, Nilsen W, Pavel M, Srivastava M. Mobile health: revolutionizing healthcare through transdisciplinary research. Comput 2013 Jan;46(1):28-35. [doi: 10.1109/MC.2012.392]
9. Lu C, Hu Y, Xie J, Fu Q, Leigh I, Governor S, et al. The use of mobile health applications to improve patient experience: cross-sectional study in Chinese public hospitals. JMIR Mhealth Uhealth 2018 May 23;6(5):e126 [FREE Full text] [doi: 10.2196/mhealth.9145] [Medline: 29792290]
10. United Nations. 2015. World Population Ageing URL: http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf [accessed 2020-05-11]
11. Ni Z, Wu B, Samples C, Shaw R. Mobile technology for health care in rural China. Int J Nurs Stud 2014 Sep;1(3):323-324 [FREE Full text] [doi: 10.1016/j.ijnss.2014.07.003]
12. Deng Z, Hong Z, Ren C, Zhang W, Xiang F. What predicts patients' adoption intention toward mhealth services in China: empirical study. JMIR Mhealth Uhealth 2018 Aug 29;6(8):e172 [FREE Full text] [doi: 10.2196/mhealth.9316] [Medline: 30158101]
13. Ye Q, Deng Z, Chen Y, Liao J, Li G, Lu Y. How resource scarcity and accessibility affect patients' usage of mobile health in China: resource competition perspective. JMIR Mhealth Uhealth 2019 Aug 9;7(8):e13491 [FREE Full text] [doi: 10.2196/13491] [Medline: 31400104]
14. Lee K, Jung SY, Hwang H, Yoo S, Baek HY, Baek R, et al. A novel concept for integrating and delivering health information using a comprehensive digital dashboard: An analysis of healthcare professionals' intention to adopt a new system and the trend of its real usage. Int J Med Inform 2017 Jan;97:98-108. [doi: 10.1016/j.ijmedinf.2016.10.001] [Medline: 27919400]
15. Kim S, Lee K, Hwang H, Yoo S. Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (EMR) using the unified theory of acceptance and use of technology (UTAUT) in a tertiary hospital. BMC Med Inform Decis Mak 2016 Jan 30;16:12 [FREE Full text] [doi: 10.1186/s12911-016-0249-8] [Medline: 26831123]
16. Dwivedi Y, Shareef M, Simintiras A, Lal B, Weerakkody V. A generalised adoption model for services: a cross-country comparison of mobile health (m-health). Gov Inf Q 2016 Jan;33(1):174-187 [FREE Full text] [doi: 10.1016/j.giq.2015.06.003]
17. Gan Q. Is the adoption of electronic health record system 'contagious'? Health Policy and Technol 2015 Jun;4(2):107-112 [FREE Full text] [doi: 10.1016/j.hlpt.2015.02.009]
18. Alaiad A, Zhou L. The determinants of home healthcare robots adoption: an empirical investigation. Int J Med Inform 2014 Nov;83(11):825-840. [doi: 10.1016/j.ijmedinf.2014.07.003] [Medline: 25132284]
19. Tsai C. Integrating social capital theory, social cognitive theory, and the technology acceptance model to explore a behavioral model of telehealth systems. Int J Environ Res Public Health 2014 May 7;11(5):4905-4925 [FREE Full text] [doi: 10.3390/ijerph110504905] [Medline: 24810577]
20. Deng Z, Mo X, Liu S. Comparison of the middle-aged and older users' adoption of mobile health services in China. Int J Med Inform 2014 Mar;83(3):210-224. [doi: 10.1016/j.ijmedinf.2013.12.002] [Medline: 24388129]
21. Phichitchaisopa N, Naenna T. Factors affecting the adoption of healthcare information technology. Excli J 2013;12:413-436 [FREE Full text] [Medline: 26417235]
22. Rai A, Chen L, Pye J, Baird A. Understanding determinants of consumer mobile health usage intentions, assimilation, and channel preferences. J Med Internet Res 2013 Aug 2;15(8):e149 [FREE Full text] [doi: 10.2196/jmir.2635] [Medline: 23912839]
23. Hung S, Ku Y, Chien J. Understanding physicians' acceptance of the medline system for practicing evidence-based medicine: a decomposed TPB model. Int J Med Inform 2012 Feb;81(2):130-142. [doi: 10.1016/j.ijmedinf.2011.09.009] [Medline: 22047627]
24. Wu L, Li J, Fu C. The adoption of mobile healthcare by hospital's professionals: an integrative perspective. Decis Support Syst 2011 Jun;51(3):587-596 [FREE Full text] [doi: 10.1016/j.dss.2011.03.003]
25. Kijsanayotin B, Pannarunothai S, Speedie SM. Factors influencing health information technology adoption in Thailand's community health centers: applying the UTAUT model. Int J Med Inform 2009 Jun;78(6):404-416. [doi: 10.1016/j.ijmedinf.2008.12.005] [Medline: 19196548]
26. Tung F, Chang S, Chou C. An extension of trust and TAM model with IDT in the adoption of the electronic logistics information system in HIS in the medical industry. Int J Med Inform 2008 May;77(5):324-335. [doi: 10.1016/j.ijmedinf.2007.06.006] [Medline: 17644029]
27. Hu P, Chau P, Sheng O, Tam K. Examining the technology acceptance model using physician acceptance of telemedicine technology. J Manage Inform Syst 2015 Dec 2;16(2):91-112 [FREE Full text] [doi: 10.1080/07421222.1999.11518247]
28. Tornatzky LG, Fleischer M. The Processes of Technological Innovation. Maryland, United States: Lexington Books; 1990.
29. Liang Y, Qi G, Wei K, Chen J. Exploring the determinant and influence mechanism of e-government cloud adoption in government agencies in China. Gov Inf Q 2017 Sep;34(3):481-495 [FREE Full text] [doi: 10.1016/j.giq.2017.06.002]

30. Lin H. Understanding the determinants of electronic supply chain management system adoption: using the technology–organization–environment framework. Technol Forecast Soc 2014 Jul;86:80-92 [FREE Full text] [doi: 10.1016/j.techfore.2013.09.001]
31. Lin H, Lin S. Determinants of e-business diffusion: a test of the technology diffusion perspective. Technovation 2008 Mar;28(3):135-145 [FREE Full text] [doi: 10.1016/j.technovation.2007.10.003]
32. Chau P, Tam K. Factors affecting the adoption of open systems: an exploratory study. MIS Q 1997 Mar;21(1):1 [FREE Full text] [doi: 10.2307/249740]
33. Kuan K, Chau P. A perception-based model for EDI adoption in small businesses using a technology–organization–environment framework. Inf Manag 2001 Oct;38(8):507-521 [FREE Full text] [doi: 10.1016/s0378-7206(01)00073-8]
34. Alam SS. Adoption of internet in Malaysian SMEs. J Small Bus Enterprise Dev 2009 May 15;16(2):240-255 [FREE Full text] [doi: 10.1108/14626000910956038]
35. Ramdani B, Chevers D, Williams D. SMEs' adoption of enterprise applications: a technology-organisation-environment model. J Small Bus Enterprise Dev 2013;20(4):735-753 [FREE Full text] [doi: 10.1108/jsbed-12-2011-0035]
36. Ramdani B, Kawalek P, Lorenzo O. Predicting SMEs' adoption of enterprise systems. J Enterprise Inf Manag 2009 Feb 13;22(1/2):10-24 [FREE Full text] [doi: 10.1108/17410390910922796]
37. Awa H, Ojiabo O. A model of adoption determinants of ERP within T-O-E framework. Inf Technol People 2016 Nov 7;29(4):901-930 [FREE Full text] [doi: 10.1108/itp-03-2015-0068]
38. Wang Y, Li H, Li C, Zhang D. Factors affecting hotels' adoption of mobile reservation systems: a technology-organization-environment framework. Tour Manag 2016 Apr;53:163-172 [FREE Full text] [doi: 10.1016/j.tourman.2015.09.021]
39. Zhu K, Kraemer K. Post-adoption variations in usage and value of e-business by organizations: cross-country evidence from the retail industry. Inf Syst Res 2005 Mar;16(1):61-84 [FREE Full text] [doi: 10.1287/isre.1050.0045]
40. Oliveira T, Thomas M, Espadanal M. Assessing the determinants of cloud computing adoption: an analysis of the manufacturing and services sectors. Inf Manag 2014 Jul;51(5):497-510 [FREE Full text] [doi: 10.1016/j.im.2014.03.006]
41. Vest JR. More than just a question of technology: factors related to hospitals' adoption and implementation of health information exchange. Int J Med Inform 2010 Dec;79(12):797-806. [doi: 10.1016/j.ijmedinf.2010.09.003] [Medline: 20889370]
42. Cao Q, Jones D, Sheng H. Contained nomadic information environments: technology, organization, and environment influences on adoption of hospital RFID patient tracking. Inf Manag 2014 Mar;51(2):225-239 [FREE Full text] [doi: 10.1016/j.im.2013.11.007]
43. Gutierrez A, Boukrami E, Lumsden R. Technological, organisational and environmental factors influencing managers' decision to adopt cloud computing in the UK. J Enterprise Inf Manag 2015 Oct 12;28(6):788-807 [FREE Full text] [doi: 10.1108/jeim-01-2015-0001]
44. Thiesse F, Staake T, Schmitt P, Fleisch E. The rise of the 'next‐generation bar code': an international RFID adoption study. Supply Chain Manag 2011 Aug 9;16(5):328-345 [FREE Full text] [doi: 10.1108/13598541111155848]
45. Wang Y, Wang Y, Yang Y. Understanding the determinants of RFID adoption in the manufacturing industry. Technol Forecast Soc 2010 Jun;77(5):803-815 [FREE Full text] [doi: 10.1016/j.techfore.2010.03.006]
46. Zhu K, Dong S, Xu S, Kraemer K. Innovation diffusion in global contexts: determinants of post-adoption digital transformation of European companies. Eur J Inform Syst 2017 Dec 19;15(6):601-616 [FREE Full text] [doi: 10.1057/palgrave.ejis.3000650]
47. Zhu K, Kraemer K, Xu S. The process of innovation assimilation by firms in different countries: a technology diffusion perspective on e-business. Manag SCi 2006 Oct;52(10):1557-1576 [FREE Full text] [doi: 10.1287/mnsc.1050.0487]
48. Gibbs J, Kraemer K. A cross-country investigation of the determinants of scope of e-commerce use: an institutional approach. Electron Mark 2004 Jun 1;14(2):124-137 [FREE Full text] [doi: 10.1080/10196780410001675077]
49. Xu S, Zhu K, Gibbs J. Global technology, local adoption: a cross-country investigation of internet adoption by companies in the United States and China. Electron Mark 2004 Apr 1;14(1):13-24 [FREE Full text] [doi: 10.1080/1019678042000175261]
50. Stewart K, Segars A. An empirical examination of the concern for information privacy instrument. Info Syst Res 2002 Mar;13(1):36-49 [FREE Full text] [doi: 10.1287/isre.13.1.36.97]
51. Jeyaraj A, Rottman J, Lacity M. A review of the predictors, linkages, and biases in IT innovation adoption research. J Info Technol 2006 Mar;21(1):1-23 [FREE Full text] [doi: 10.1057/palgrave.jit.2000056]
52. Venkatesh V, Brown S, Maruping L, Bala H. Predicting different conceptualizations of system use: the competing roles of behavioral intention, facilitating conditions, and behavioral expectation. MIS Q 2008;32(3):483-502 [FREE Full text] [doi: 10.2307/25148853]
53. Sligo J, Gauld R, Roberts V, Villa L. A literature review for large-scale health information system project planning, implementation and evaluation. Int J Med Inform 2017 Jan;97:86-97. [doi: 10.1016/j.ijmedinf.2016.09.007] [Medline: 27919399]

54. Goodhue D, Thompson R. Task-technology fit and individual performance. MIS Q 1995 Jun;19(2):213-236 [FREE Full text] [doi: 10.2307/249689]
55. Lin A, Chen N. Cloud computing as an innovation: percepetion, attitude, and adoption. Int J Info Manag 2012 Dec;32(6):533-540 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2012.04.001]
56. Pagani M. Determinants of adoption of high speed data services in the business market: evidence for a combined technology acceptance model with task technology fit model. Inf Manag 2006 Oct;43(7):847-860 [FREE Full text] [doi: 10.1016/j.im.2006.08.003]
57. Ministry OH. Guiding principles for grading nursing in general hospitals (for trial implementation). China Nurs Manag 2009;9(6):33-34.
58. Chen H, Bates RA. Instrument Translation and Development Strategies for Cross-Cultural Studies, in Proceedings of the Academy of Human Resource Development 2005;693?700. 2019 Presented at: -; -; -.
59. Moore G, Benbasat I. Development of an instrument to measure the perceptions of adopting an information technology innovation. Inf Syst Res 1991 Sep;2(3):192-222 [FREE Full text] [doi: 10.1287/isre.2.3.192]
60. Kim D, Ferrin D, Rao H. A trust-based consumer decision-making model in electronic commerce: the role of trust, perceived risk, and their antecedents. Decis Support Syst 2008 Jan;44(2):544-564 [FREE Full text] [doi: 10.1016/j.dss.2007.07.001]
61. Bhattacherjee A, Hikmet N. Reconceptualizing organizational support and its effect on information technology usage: evidence from the health care sector. J Compt Info Syst 2008;48(4):69-76 [FREE Full text] [doi: 10.1080/08874417.2008.11646036]
62. Yap C, Thong J, Raman K. Effect of government incentives on computerisation in small business. Eur J Inform Syst 2017 Dec 19;3(3):191-206 [FREE Full text] [doi: 10.1057/ejis.1994.20]
63. Grandon E, Pearson J. Electronic commerce adoption: an empirical study of small and medium US businesses. Ing Manag 2004 Dec;42(1):197-216 [FREE Full text] [doi: 10.1016/j.im.2003.12.010]
64. Premkumar G, Roberts M. Adoption of new information technologies in rural small businesses. Omega 1999 Aug;27(4):467-484 [FREE Full text] [doi: 10.1016/s0305-0483(98)00071-1]
65. Hair J, Black W, Babin B, Anderson R, Tatham R. Multivariate Data Analysis: With Readings. Edinburgh, UK: Pearson Education; 2010.
66. Shareef M, Kumar V, Kumar U. Predicting mobile health adoption behaviour: a demand side perspective. J Cust Behav 2014 Oct 31;13(3):187-205 [FREE Full text] [doi: 10.1362/147539214x14103453768697]
67. Wu J, Wang S, Lin L. Mobile computing acceptance factors in the healthcare industry: a structural equation model. Int J Med Inform 2007 Jan;76(1):66-77. [doi: 10.1016/j.ijmedinf.2006.06.006] [Medline: 16901749]
68. Villalba-Mora E, Casas I, Lupiañez-Villanueva F, Maghiros I. Adoption of health information technologies by physicians for clinical practice: the Andalusian case. Int J Med Inform 2015 Jul;84(7):477-485. [doi: 10.1016/j.ijmedinf.2015.03.002] [Medline: 25823578]
69. Avgar AC, Litwin AS, Pronovost PJ. Drivers and barriers in health IT adoption: a proposed framework. Appl Clin Inform 2012;3(4):488-500 [FREE Full text] [doi: 10.4338/ACI-2012-07-R-0029] [Medline: 23646093]
70. Vishwanath A, Scamurra SD. Barriers to the adoption of electronic health records: using concept mapping to develop a comprehensive empirical model. Health Informatics J 2007 Jun;13(2):119-134. [doi: 10.1177/1460458207076468] [Medline: 17510224]

Abbreviations

CRM: customer relationship management
e-business: electronic business
HIE: health information exchange
IT: information technology
mHealth: mobile health
RFID: radio-frequency identification
SME: small-to-medium–sized enterprise
TMS: top management support
TOE: technology-organization-environment

Edited by C Lovis; submitted 23.05.19; peer-reviewed by W Zhang, K Ahmed, Z Li; comments to author 07.08.19; revised version received 18.11.19; accepted 22.11.19; published 14.07.20.

Please cite as:
Ramdani B, Duan B, Berrou I
Exploring the Determinants of Mobile Health Adoption by Hospitals in China: Empirical Study
JMIR Med Inform 2020;8(7):e14795
URL: https://medinform.jmir.org/2020/7/e14795
doi:10.2196/14795
PMID:32459630

Original Paper

Extraction of Information Related to Drug Safety Surveillance From Electronic Health Record Notes: Joint Modeling of Entities and Relations Using Knowledge-Aware Neural Attentive Models

Bharath Dandala[1], BTECH, MS, PhD; Venkata Joopudi[1], BTECH, MS; Ching-Huei Tsou[1], BS, MEng, PhD; Jennifer J Liang[1], SB, MD; Parthasarathy Suryanarayanan[1], BSc, BTECH

IBM Research, Yorktown Heights, NY, United States

Corresponding Author:
Bharath Dandala, BTECH, MS, PhD
IBM Research
1101 Kitchawan Rd
Yorktown Heights, NY, 10598
United States
Phone: 1 9403673972
Email: bdand@us.ibm.com

Abstract

Background: An adverse drug event (ADE) is commonly defined as "an injury resulting from medical intervention related to a drug." Providing information related to ADEs and alerting caregivers at the point of care can reduce the risk of prescription and diagnostic errors and improve health outcomes. ADEs captured in structured data in electronic health records (EHRs) as either coded problems or allergies are often incomplete, leading to underreporting. Therefore, it is important to develop capabilities to process unstructured EHR data in the form of clinical notes, which contain a richer documentation of a patient's ADE. Several natural language processing (NLP) systems have been proposed to automatically extract information related to ADEs. However, the results from these systems showed that significant improvement is still required for the automatic extraction of ADEs from clinical notes.

Objective: This study aims to improve the automatic extraction of ADEs and related information such as drugs, their attributes, and reason for administration from the clinical notes of patients.

Methods: This research was conducted using discharge summaries from the Medical Information Mart for Intensive Care III (MIMIC-III) database obtained through the 2018 National NLP Clinical Challenges (n2c2) annotated with drugs, drug attributes (ie, strength, form, frequency, route, dosage, duration), ADEs, reasons, and relations between drugs and other entities. We developed a deep learning–based system for extracting these drug-centric concepts and relations simultaneously using a joint method enhanced with contextualized embeddings, a position-attention mechanism, and knowledge representations. The joint method generated different sentence representations for each drug, which were then used to extract related concepts and relations simultaneously. Contextualized representations trained on the MIMIC-III database were used to capture context-sensitive meanings of words. The position-attention mechanism amplified the benefits of the joint method by generating sentence representations that capture long-distance relations. Knowledge representations were obtained from graph embeddings created using the US Food and Drug Administration Adverse Event Reporting System database to improve relation extraction, especially when contextual clues were insufficient.

Results: Our system achieved new state-of-the-art results on the n2c2 data set, with significant improvements in recognizing crucial drug−reason (F1=0.650 versus F1=0.579) and drug−ADE (F1=0.490 versus F1=0.476) relations.

Conclusions: This study presents a system for extracting drug-centric concepts and relations that outperformed current state-of-the-art results and shows that contextualized embeddings, position-attention mechanisms, and knowledge graph embeddings effectively improve deep learning–based concepts and relation extraction. This study demonstrates the potential for deep learning–based methods to help extract real-world evidence from unstructured patient data for drug safety surveillance.

(JMIR Med Inform 2020;8(7):e18417) doi:10.2196/18417

KEYWORDS

electronic health records; adverse drug events; natural language processing; deep learning; information extraction; adverse drug reaction reporting systems; named entity recognition; relation extraction

Introduction

Background

An electronic health record (EHR) is the systematized collection of electronically stored health information of patients and the general population in a digital format [1]. Clinical notes in EHRs summarize interactions that occur between patients and health care providers [2]. These notes include observations, impressions, treatments, drug use, adverse drug events (ADEs), and other activities arising from each interaction between the patient and the health care system. Extracting useful information such as ADEs from these notes and alerting caregivers at the point of care has the potential to improve patient health outcomes.

An ADE is commonly defined as "an injury resulting from medical intervention related to a drug" [3]. ADEs are a major public health concern and one of the leading causes of morbidity and mortality [4]. Studies have shown the substantial economic burden of these undesired effects [5,6]. Although drug safety and efficacy are tested during premarketing randomized clinical trials, these trials may not detect all ADEs because such studies are often small, short, and biased by the exclusion of patients with comorbid diseases. With the limited information available when a drug is marketed, postmarketing surveillance has become increasingly important. Spontaneous reporting systems, such as the US Food and Drug Administration Adverse Event Reporting System (FAERS) [7], are monitoring mechanisms for postmarketing surveillance that enable both physicians and patients to report ADEs. However, previous studies [8-10] have exposed various inadequacies with such systems, including underreporting, reporting biases, and incomplete information, prompting researchers to explore additional sources to detect ADEs from real-world data.

Several efforts have been made to extract ADEs automatically from disparate information sources, including EHRs [11-13], spontaneous reporting systems [14-16], social media [17-20], search queries on the web via search engine logs [21,22], and biology and chemistry knowledge bases [23-25]. Furthermore, the clinical natural language processing (NLP) community has organized several open challenges such as the 2010 Informatics for Integrating Biology & the Bedside/Veterans Affairs NLP Challenge [26], Text Analysis Conference 2017 Adverse Drug Reactions Track [27], and BioCreative V Chemical Disease Relation task [28]. Recently, 2 such challenges, Medication and Adverse Drug Events from Electronic Health Records (MADE 1.0) [29] and the 2018 National NLP Clinical Challenges (n2c2) Shared Task Track 2 [30], were organized to extract *drugs*, drug attributes, *ADEs*, *reasons* for prescribing drugs, and their relations from clinical notes. The results from these 2 challenges showed that deep learning techniques outperform traditional machine learning techniques for this task, and significant improvement is still required for *drug−{ADE, reason}* relation extraction. Specifically, the organizers of these challenges hypothesized that models that can effectively incorporate the larger context to capture long-distance relations or leverage knowledge to capture implicit relations will likely improve the performance of future systems.

Considering these conclusions, we developed a joint deep learning–based relation extraction system that helps in extracting long-distance relations through a position-attention mechanism and implicit relations through external knowledge from the FAERS. To the best of our knowledge, no previous research has been conducted on using the position-attention mechanism and domain-specific knowledge graph embeddings in ADE detection.

Relevant Literature

Adverse Drug Event Detection

From the viewpoint of NLP, effective techniques for entity and relation extraction are fundamental requirements in automatic ADE extraction. Entity and relation extraction from text has traditionally been treated as a pipeline of 2 separate subtasks: named entity recognition (NER) and relation classification. Previous studies employed traditional machine learning techniques [31-34], such as conditional random fields (CRF) [35] for NER and support vector machines [36] for relation classification. Several recent approaches [37-44], developed on MADE 1.0 [29] and 2018 n2c2 Shared Task Track 2 [30] data sets, employed deep learning techniques, such as bidirectional, long short-term memory–conditional random fields (BiLSTM-CRFs) [45], for NER and convolutional neural network (CNN) [46] for relation classification, and showed numerous advantages resulting in better performance and less feature engineering. However, there is an inevitable error propagation issue with pipeline-based methods because of the following:

1. NER relying on sequence-labeling techniques suffers from lossy representation when there are overlapping annotations on entities. For example, in "she was on *furosemide* and became *hypotensive* requiring *norepinephrine*," *hypotensive* is an *ADE* with respect to *furosemide* but a *reason* with respect to *norepinephrine*.
2. NER approaches usually take an input context window that may not contain the necessary information to determine the appropriate label (ie, *ADE*, *reason*, no label). For example, in "Patient reports *nausea*. Started on *ondansetron*," the identification of *nausea* as a *reason* requires information from both sentences.
3. Signs or symptoms are only labeled as *ADE* or *reason* if they are related to a drug (ie, not all signs or symptoms in the clinical note are annotated). This makes the corpus less suitable to train an effective relation classification model as it misses negative candidate pairs for *drug−{ADE, reason}* relations.

To address the first 2 issues, we previously proposed a joint method that outperformed the pipeline method for concept and

relation extraction on a similar data set (MADE 1.0) [37]. In a separate study, Li et al [47] proposed a joint method using multitask learning [48] and made similar observations. To address the third issue, which was introduced with the n2c2 data set, Wei et al [38] proposed a novel label-encoding scheme to jointly extract *ADE*, *reason*, drug attributes, and their relations.

Attention-Based Relation Extraction

The attention mechanism allows neural networks to selectively focus on specific information [49-51]. This has proven to be effective for NLP problems with long-distance dependencies such as NER and relation extraction. Zhou et al [52] proposed an attention-based BiLSTM network and demonstrated its effectiveness in selectively focusing on words that have decisive effects on relation classification. Next, Zhang et al [53] extended the attention mechanism to help networks not only focus on words based on the semantic information of the sentence but also the global positions of entities within the sentence. Recently Dai et al [54] introduced a position-attention mechanism for joint extraction of entities and overlapping relations. The position-attention mechanism builds on self-attention by focusing on both the global dependencies of the input and tokens of the target entities of interest for relation extraction. Recent research [37,55] on ADE extraction showed the benefits of self-attention mechanisms in pipeline-based methods, specifically for relation classification. However, to the best of our knowledge, no previous work has focused on using self-attention or position-attention mechanisms for joint extraction of entities and relations for ADE extraction.

Knowledge-Aware Relation Extraction

Several approaches [56-59] in the open domain have shown that incorporating embeddings learned from knowledge bases benefit deep learning–based relation classification. These embeddings are typically learned using translation-based methods such as TransE [60], TransH [61], and TransR [62]; walk-based methods such as DeepWalk [63] and node2vec [64]; or neural network–based methods such as large-scale information network embedding (LINE) [65] and bipartite network embedding [66].

Clinical notes are typically written for medical professionals. Hence, a certain degree of medical knowledge is assumed by the authors, which is not explicitly expressed in the text. This is especially true for relations between clinical findings and drugs, where a drug could either cause (*ADE*) or treat (*reason*) a clinical finding. In our previous study [37], we showed that augmenting knowledge base features such as proportional report ratio and reporting odds ratio calculated from the FAERS into deep learning models can benefit relation classification. Recently, Chen et al [67] proposed a hybrid clinical NLP system by combining a general knowledge-based system using the Unified Medical Language System (UMLS) and BiLSTM-CRF for concept extraction and attention-BiLSTM for relation classification. However, to the best of our knowledge, no previous work has focused on using knowledge graph embeddings generated from the FAERS for joint extraction of entities and relations for ADE extraction.

Methods

Data Set

The n2c2 data set consists of 505 deidentified clinical narratives, of which 303 and 202 narratives were released as train and test data sets, respectively. Each narrative was manually annotated with drug-centric entities, including *drugs*, their attributes (*strength*, *form*, *frequency*, *route*, *dosage*, and *duration*), *ADEs*, *reasons*, and relations between drugs and other entities (*drug*−{attributes, *ADE*, *reason*}). *Drug*−{attributes} represent 6 different types of relations: *drug*−{*strength*, *form*, *frequency*, *route*, *dosage*, *duration*}. Figure 1 presents an example with annotations. Tables 1 and 2 present the statistical overview of the annotated entities and relations.

Figure 1. An illustration with annotations for entities and relations. ADE: adverse drug event; HTN: hypertension; QHS: every night at bedtime.

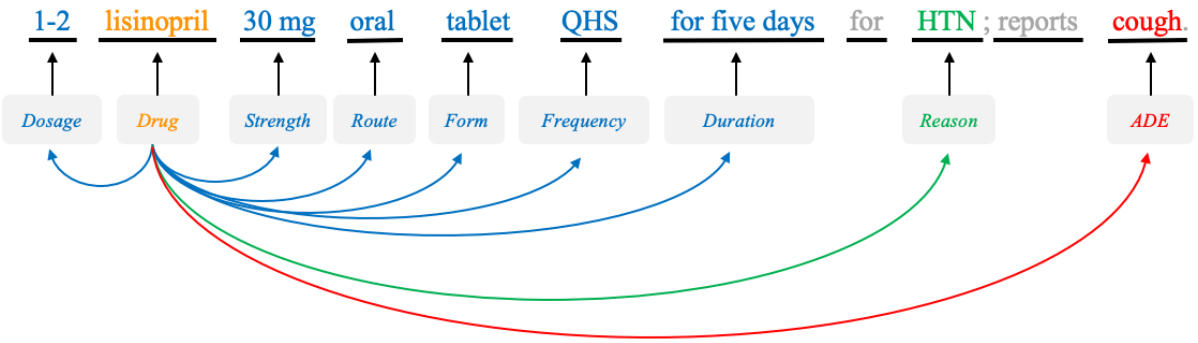


Table 1. Entities in the data set.

Entity type	Number of annotations		Example	Description
	Train, n (%)	Test, n (%)		
Drug	16,225 (31.84)	10,575 (32.13)	Coumadin	Name of the drug
Strength	6691 (13.13)	4230 (12.85)	5 mg	Strength of the drug
Form	6651 (13.05)	4359 (13.24)	Tablet	Form of the drug
Frequency	6281 (12.32)	4012 (12.19)	Daily	Frequency of the drug
Route	5476 (10.75)	3513 (10.67)	By mouth	Route in which the drug is administered
Dosage	4221 (8.28)	2681 (8.14)	1	Dosage of the drug
Duration	592 (1.16)	378 (1.15)	For 5 days	Duration of the drug
ADE[a]	959 (1.88)	625 (1.90)	Rash	Adverse reaction of the drug
Reason	3855 (7.57)	2545 (7.73)	Constipation	Indication if it is an affliction that a physician is actively treating with a drug
Total	50,951 (100.00)	32,918 (100.00)	N/A[b]	N/A

[a]ADE: adverse drug event.

[b]Not applicable.

Table 2. Relations in the data set.

Relation type	Relations		Intersentential relations		Example[a]
	Train, n (%)	Test, n (%)	Train, n (%)	Test, n (%)	
Drug−strength	6702 (18.44)	4244 (18.09)	80 (1.19)	59 (1.39)	*Lisinopril* 1×*5 mg* tablet orally daily for 7 days
Drug−form	6654 (18.31)	4374 (18.64)	259 (3.89)	144 (3.29)	*Lisinopril* 1×5 mg *tablet* orally daily for 7 days
Drug−frequency	6310 (17.36)	4034 (17.19)	372 (5.90)	238 (5.90)	*Lisinopril* 1×5 mg tablet orally *daily* for 7 days
Drug−route	5538 (15.24)	3546 (15.11)	199 (3.59)	149 (4.20)	*Lisinopril* 1×5 mg tablet *orally* daily for 7 days
Drug−dosage	4225 (11.62)	2695 (11.49)	135 (3.20)	102 (3.78)	*Lisinopril1*×5 mg tablet orally daily for 7 days
Drug−duration	643 (1.80)	426 (1.80)	34 (5.4)	43 (10.0)	*Lisinopril* 1×5 mg tablet orally daily for *7 days*
Drug−ADE[b]	1107 (3.05)	733 (3.10)	254 (22.94)	139 (18.9)	Patient is experiencing *muscle pain*, secondary to *statin* therapy for coronary artery disease
Drug−reason	5169 (14.22)	3410 (14.53)	1638 (31.69)	1088 (31.91)	Patient is experiencing muscle pain, secondary to *statin* therapy for *coronary artery disease*
Total	36,348 (100.00)	23,462 (100.00)	2971 (8.17)	1947 (8.30)	N/A[c]

[a]Italics indicate entities participating in the specified relation type.

[b]ADE: adverse drug event.

[c]Not applicable.

Preprocessing

Sentence boundary detection (SBD) and tokenization are often treated as solved problems in NLP and carried out using off-the-shelf toolkits such as Apache Natural Language Toolkit [68], Explosion AI spaCy [69] or the Stanford CoreNLP toolkit [70]. However, these are still difficult and critical problems [71] in the clinical domain because (1) sentence ends are frequently indicated by layout and not by punctuation and (2) white space is not always present to indicate token boundaries (eg, *50 mg*). To address these issues, we incorporated domain-specific rules sensitive to low-level features such as capitalization, text-wrap properties, indentation, and punctuation into the spaCy tokenizer and SBD models. These custom rules are provided in Multimedia Appendix 1.

Representation Learning

Static Word Representations

Word embedding is a text vectorization technique that transforms words or subwords into vectors of real numbers. Pretrained word embeddings created using Word2Vec [72], Glove [73], and fastText [74] have been broadly used to initialize deep learning architectures for NLP tasks and have shown substantial improvement over random initialization. Recent research [75] showed that NER performance is significantly affected by the overlap between the pretrained

word embedding vocabulary and the vocabulary of the target NER data set. Thus, we used Word2Vec with skip-gram to pretrain word embeddings over the Medical Information Mart for Intensive Care III (MIMIC-III) [76] with the default parameters provided in a study by Mikolov et al [72].

Contextualized Word Representations

A well-known limitation of word embedding methods is that they produce a single representation of all possible meanings of a word. To tackle these deficiencies, advanced approaches have attempted to model the word's context into a vector representation. Embeddings from Language Models (ELMo) [77] is a prominent model that generates contextualized word representations by combining the internal states of different layers in a neural language model. Bidirectional Enconder Representations from Transformers (BERT) [78] furthered this idea by training bidirectional transformers [50] using subwords. Contextualized embeddings are particularly useful for clinical NER as entities (eg, *cold* as low temperature versus infection) have different meanings in different contexts. Recent research [79] showed that deep learning architectures with contextualized embeddings pretrained on a large clinical corpus achieve state-of-the-art performance on several clinical NER data sets. Inspired by these, we trained contextualized representations using ELMo on MIMIC-III. Detailed explanations of ELMo and training parameters are provided in Multimedia Appendix 2.

Knowledge Representations

To introduce medical knowledge, we built knowledge representations on the FAERS, a database for postmarketing drug safety monitoring. Specifically, we used 2 tables from Adverse Event Open Learning through Universal Standardization (AEOLUS) [14], a curated and standardized FAERS resource, to generate 2 separate graph embeddings. As shown in Figure 2, *standard drug_outcome count* contains case frequencies for drug outcomes, including *ADEs*, and *standard drug indication count* contains case frequencies for drug indications (ie, *reasons)*.

Let $G=(D,O,E)$ be a weighted bipartite network, where D and O denote the set of *drug concept id* and *outcome concept id* in *standard drug outcome count*, and E defines the interset edges. D_i and O_j denote the i^{th} and j^{th} vertex in D and O respectively, where $i=\{1,2,\ldots,|D|\}$ and $j=\{1,2,\ldots,|O|\}$. Each edge e_{ij} carries a frequency f_{ij} provided by the *drug outcome pair count* field in *standard drug outcome count*, indicating the strength between the connected vertices D_i and O_j; if D_i and O_j are not connected, f_{ij} is set to zero. To integrate this knowledge into our proposed architecture, we computed token-level embeddings by transforming G to G' as follows:

Given a *drug concept id* (RxNorm) or *outcome concept id* (Medical Dictionary for Regulatory Activities) from AEOLUS, we mapped it to its concept unique identifiers (CUIs) in UMLS [80] and obtained a set of tokens from all CUI variants. Let $d=\{d_1, d_2, \ldots, d_L\}$ and $o=\{o_1, o_2, \ldots, o_M\}$ represent all unique drug and outcome tokens obtained from mapping all D_i and O_j. Let ϕ_d and ϕ_o represent 2 multivalued functions that associate each element in the set of *drug concept id* and *outcome concept id* to a set of tokens. Let $G'=(d,o,e)$ be a weighted bipartite graph and each edge e_{ij} of G' is associated with a nonnegative weight w_{lm} indicating the strength between the drug token d_l and the outcome token o_m. We calculated w_{lm} as token-level co-occurrence between d_l and o_m normalized for the drug token d_l:

$$w_{lm} = \frac{\sum_{i,j} f_{ij} \cdot \mathbb{1}_{d_l \in \phi_d(D_i)} \cdot \mathbb{1}_{o_m \in \phi_o(O_j)}}{\sum_{i,j} f_{ij} \cdot \mathbb{1}_{d_l \in \phi_d(D_i)}}$$

In w_{lm}, the numerator represents the sum of frequencies of all *drug concept id* and *outcome concept id* pairs that contain drug token d_l and outcome token o_m and the denominator represents the sum of frequencies of all pairs whose *drug concept id* contains the drug token d_l.

From the generated bipartite weighted graph $G'=(d,o,e)$, we used the LINE approach to generate *drug-adverse* knowledge embeddings. We used LINE because (1) relations between drugs and other concepts in the FAERS form a weighted bipartite graph with a long-tail distribution of vertex degrees and (2) it helps in embedding implicit connectivity relations between vertices of the same type. Similarly, we generated *drug-reason* knowledge embeddings from the *standard drug indication count* table. Detailed explanations of LINE and training parameters are provided in Multimedia Appendix 2.

Figure 2. Excerpts from the standard drug outcome count and standard drug indication count tables from adverse event open learning through universal standardization.

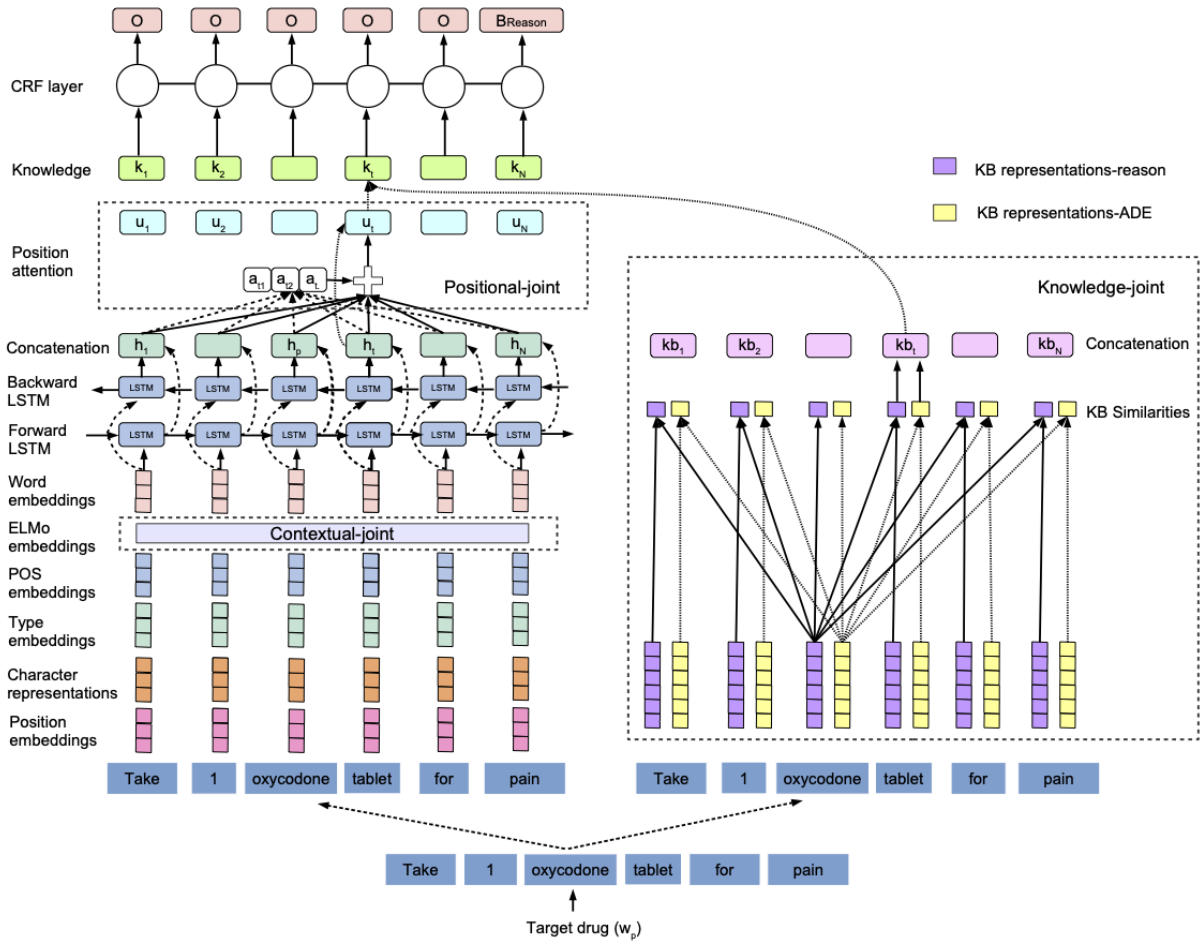
standard_drug_outcome_count				
drug_concept_id	Drug name	outcome_concept_id	outcome name	drug_outcome_pair_count
29046	Lisinopril	10011224	Cough	103
6809	Metformin	10047700	Vomiting	399

standard_drug_indication_count				
drug_concept_id	Drug name	indication_concept_id	indication name	drug_indication_pair_count
29046	Lisinopril	10020772	Hypertension	9003
6809	Metformin	10012601	Diabetes Mellitus	9370

Architecture

In the following sections, we present our system, illustrated in Figure 3, in an incremental fashion: *joint method*, *contextual-joint*, *positional-joint*, and *knowledge-joint*. A detailed explanation of the deep learning architecture, BiLSTM-CRF [81], and input embeddings used in this system is included in the Multimedia Appendix 3.

Figure 3. Canonical architecture of the proposed system. ADE: adverse drug event; BReason: beginning of reason annotation; CRF: conditional random field; ELMo: Embeddings from Language Models; KB: knowledge base; LSTM: long short-term memory; POS: part-of-speech.



Joint Method

We developed a *drug recognition model* followed by 2 joint *drug-centric relation extraction models* (*drug*−{attributes} and *drug*−{*ADE*, *reason*}), as explained in the following sections.

Drug Recognition Model

We modeled drug recognition as a sequence-labeling task using BiLSTM-CRF and a beginning, inside, and outside of a drug mention (BIO) tagging scheme. The input layer of the BiLSTM-CRF takes word, character, and part-of-speech embeddings. The word embeddings were obtained using Word2Vec representations generated using MIMIC-III. The character and part-of-speech embeddings were initialized randomly. We used CNNs [46] to encode a character-level representation for a word.

Drug-Centric Relation Extraction Models

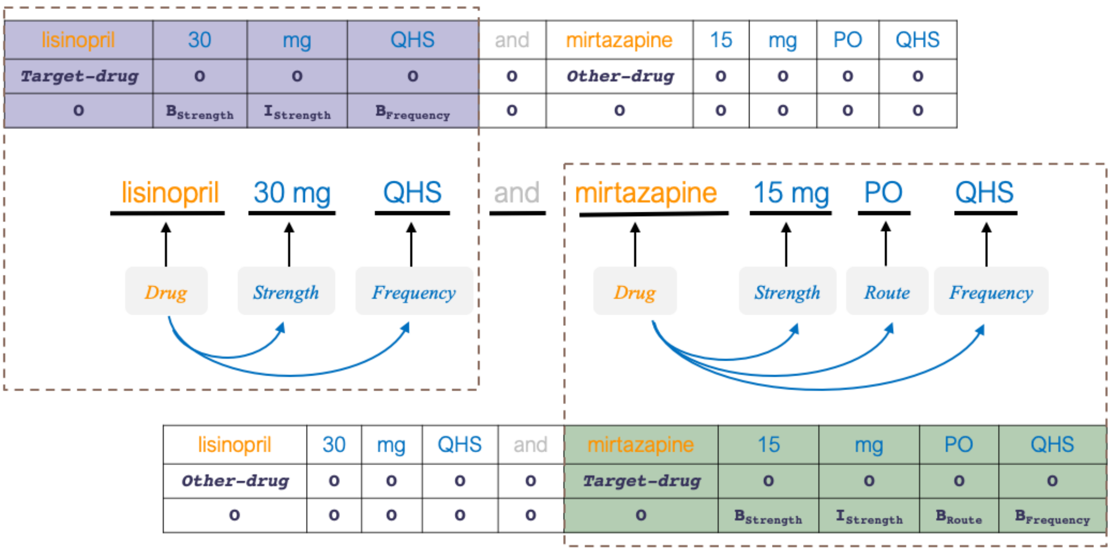
To extract entities and relations jointly, we used the encoding scheme proposed in [38], which takes annotated sentences and produces drug-centric sequences for a specified *target-drug*. For sentences containing multiple identified drugs, 1 drug-centric sequence was generated for each *target-drug*. For example, for the sentence in Figure 4, the encoding scheme produced 2 labeled sequences: one with *lisinopril* as the *target-drug* and the other with *mirtazapine*. In each sequence, associated entities with the *target-drug* were labeled using a BIO scheme enhanced with their types. Hence, for the sequence generated with *lisinopril* as the *target-drug*, only *30 mg* and the first *QHS* were labeled using B and I tags, and other entities (eg, *15 mg*, *PO*, and the second *QHS*) were labeled as *O*.

We trained 2 separate models with the BiLSTM-CRF to jointly recognize (1) drug attributes and *drug*−{attributes} relations and (2) *ADE*, *reason*, and their corresponding relations (*drug*−{*ADE*, *reason*}). Similar to the *drug recognition model*, the input layer of these models takes word, character, and part-of-speech representations, with additional positional and semantic-tag embeddings. We used the positional embedding technique introduced in [82] to represent the positional distance

from *target-drug* to each word in the input context. We used 3 different semantic tags, *target-drug*, *duplicate-target-drug*, and *nontarget-drugs*, to represent tokens of the current *target-drug*, other mentions of the same *target-drug*, and other drugs in the input context, respectively.

To handle intersentential relations, we provided adjacent sentences as an input context to the sentence containing the *target-drug*. We used training data to determine the optimal input context for the 2 models empirically. For the *drug*−{attributes} model, we determined the optimal context as the current sentence with the *target-drug* and the sentences preceding and following it. For the *drug*−{*ADE*, *reason*} model, the optimal context was the current sentence and the 4 sentences preceding and following it.

Figure 4. Label-encoding scheme used in drug-centric relation extraction models. B: beginning; I: inside; PO: orally; QHS: every night at bedtime.



Contextual-Joint Model

We obtained domain-specific contextualized representations for input contexts by pretraining ELMo on MIMIC-III. These contextualized representations were used to augment the representations used in the input layers of the models in the *joint method*. With the augmented input representations, we trained (1) a *drug recognition model* and (2) 2 *drug-centric relation extraction models* (*drug*−{attributes} and *drug*−{*ADE*, *reason*}).

Positional-Joint Model

As the task involves extraction of drug-centric entities and relations, we used the position-attention mechanism to extract entities and relations jointly with respect to an entity of interest (*target-drug*).

Let represent the hidden representations of an input sequence obtained from the BiLSTM layer of the *contextual-joint model*.

Positional representations were generated as follows:

where v, W^p, W^t, W^j are parameters to be learned, and s_{tj} is the score obtained through additive attention. Position-attention computes dependencies among the hidden states: (1) h_p at *target-drug* position p, (2) h_j at j^{th} token in the input sequence, and (3) h_t at current token t. For each token j, s_{tj} is computed by (1) comparing h_p with h_j and (2) comparing h_t with h_j The comparison of h_p and h_j helps to encode *target-drug* (positional) information, whereas the comparison of h_t and h_j is useful for matching sentence representations against itself (self-matching) to collect contextual information. a_{tj} is the attention weight produced by the normalization of s_{tj} and is used in computing the positional representation p_t of the current token t. Finally, we concatenated this positional representation p_t with its hidden representation h_t to obtain u_t:

We trained the 2 *drug-centric relation extraction models* (*drug*−{attributes} and *drug*−{*ADE*, *reason*}) by feeding these concatenated representations to a CRF layer. During the test phase, we used the *drug recognition model* from the *contextual-joint* for predicting *drugs* and the trained *drug-centric relation extraction models* for predicting *drug*−{attributes} and *drug*−{*ADE*, *reason*} relations.

Knowledge-Joint Model

As introduced earlier, background knowledge and hidden relations beyond the contextual and positional information play

a crucial role in extracting *drug−{ADE, reason}* relations. To address this, we propose the *knowledge-joint* model by enhancing the *positional-joint* model with knowledge embeddings created using the FAERS database.

Let  ,  denote representations of the input sequence tokens obtained from the *drug-reason* and *drug-adverse* knowledge embeddings, respectively. Let *l* and *m* be the beginning and end indices of *target-drug* in the input sequence. The *target-drug* D_r and D_a, corresponding to *drug-reason* and *drug-adverse* knowledge embeddings, were computed by averaging the representations of *target-drug* tokens:

The *target-drug*–centric representations  and  were obtained by computing similarities between input sequence tokens and the *target-drug*:

where w_r and w_a represent the scalar weights corresponding to *drug-reason*, and *drug-adverse* knowledge embeddings learned during training. Finally, for a token at position *t*, we concatenated its *target-drug*–centric similarities  with positional and hidden representations u_t to produce k_t:

We trained a *drug-centric relation extraction model* (*drug−{ADE, reason}*) by feeding these concatenated representations to a CRF layer. During the test phase, we used the *drug recognition model* from the *contextual-joint* model for predicting *drugs* and the trained *drug−{ADE, reason} model* for predicting *drug−ADE* and *drug−reason* relations.

Evaluation Metrics and Significance Tests

We evaluated the proposed system using the evaluation script released by the organizers of the n2c2 challenge to measure the lenient precision, recall, and F_1 scores, explained as follows. For NER, a predicted entity is considered as a true-positive if its span overlaps with a gold annotation and is the correct entity type. For relation extraction, a predicted relation is considered as a true-positive if both entities in the relation are true-positives and the relation type matches the gold annotation. We also report statistical significance on these results with 50,000 shuffles and a significance level set to .05 by using a test script released by the n2c2 organizers based on the approximate randomization test [83].

In the following sections, we present the results of our system. The experimental settings used to achieve these results are provided in Multimedia Appendix 4.

Results

Named Entity Recognition

Table 3 presents the results for each proposed incremental approach for NER. Compared with the *joint method*, incorporating contextualized embeddings (*contextual-joint model*) improved the overall microaveraged F_1 score by 0.3 percentage points. The improvement was mainly observed in recognizing *drugs* (0.6 points), with some improvements in recognizing *strength* and *reason*. Compared with the *contextual-joint model*, the *positional-joint model* improved the overall micro-F_1 score by 0.2 points, with significant improvements observed in identifying *reason* (2.1 points) and *ADE* (6.8 points). Compared with the *positional-joint model*, the *knowledge-joint model* further improved the overall micro-F_1 score by 0.1 points, with significant improvements observed in accurately determining *reason* (1.9 points) and *ADE* (1.7 points). Note that the overall improvement between the *positional-joint* and *knowledge-joint models* is relatively small due to the biased distribution of annotations, as *ADE* and *reason* together constitute less than 10% of the entities.

Significance tests showed that the improvements in micro-F_1 score observed with each incremental approach are statistically significant with *P* values of .001, <.001, and <.001 for the *contextual-joint*, *positional-joint*, and *knowledge-joint* models, respectively. As the *contextual-joint* and *positional-joint models* share the same *drug recognition model*, we ignored drug predictions when performing significance tests. Similarly, the *positional-joint* and *knowledge-joint* models share the same *drug recognition model* and *drug−{attributes} model;* therefore, we considered only *ADE* and *reason* predictions when performing significance tests.

Table 3. Lenient precision, recall, and F1 score of the proposed approaches for named entity recognition.

Entity type	Joint			Contextual-joint			Positional-joint			Knowledge-joint		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Drug	0.956	0.952	0.954	0.956	0.964	0.960	0.956	0.964	0.960	0.956	0.964	0.960
Strength	0.980	0.969	0.974	0.982	0.971	0.976	0.985	0.976	0.980	0.985	0.976	0.980
Form	0.974	0.942	0.958	0.975	0.939	0.957	0.972	0.943	0.958	0.972	0.943	0.958
Frequency	0.981	0.958	0.970	0.981	0.958	0.969	0.979	0.964	0.971	0.979	0.964	0.971
Route	0.964	0.942	0.953	0.962	0.943	0.952	0.950	0.949	0.949	0.950	0.949	0.949
Dosage	0.943	0.938	0.941	0.941	0.937	0.939	0.936	0.957	0.946	0.936	0.957	0.946
Duration	0.887	0.788	0.835	0.914	0.791	0.848	0.880	0.815	0.846	0.880	0.815	0.846
ADE[a]	0.649	0.358	0.462	0.643	0.346	0.450	0.660	0.426	0.518	0.589	0.490	0.535
Reason	0.757	0.611	0.676	0.747	0.636	0.687	0.747	0.672	0.708	0.753	0.702	0.727
Overall (micro)	0.948	0.912	0.929	0.947	0.917	0.932	0.943	0.926	0.934	0.941	0.930	0.935

[a]ADE: adverse drug event.

Relation Extraction

Table 4 presents the results for each proposed incremental approach for relation extraction. Compared with the *joint method*, the *contextual-joint* model improved the overall micro-F1 score by 0.5 percentage points, with the majority of improvements observed in accurately recognizing *drug−strength*, *drug−frequency*, *drug−reason*, and *drug−dosage relations*. Compared with the *contextual-joint model*, the *positional-joint model* improved the F_1 score by 0.4 points with significant improvements observed in determining *drug−ADE* (5.6 points) and *drug−reason* (2.9 points) relations. *The knowledge-joint model* further improved the overall F_1 score by 0.1 points, with specific improvements in *drug−ADE* by 3.0 points and *drug−reason* by 1.7 points when compared with the *positional-joint model*. Similar to the NER significance results, significance testing for relation extraction showed that the improvements observed with each incremental approach are statistically significant with *P* values of <.001, <.001, and <.001 for the *contextual-joint*, *positional-joint*, and *knowledge-joint* models, respectively.

Table 4. Lenient precision, recall, and F1 score of the proposed approaches for relation extraction.

Relation type	Joint			Contextual-joint			Positional-joint			Knowledge-joint		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Drug−strength	0.966	0.962	0.964	0.977	0.964	0.971	0.978	0.971	0.975	0.978	0.971	0.975
Drug−form	0.963	0.936	0.949	0.972	0.936	0.953	0.969	0.939	0.954	0.969	0.939	0.954
Drug−frequency	0.961	0.949	0.955	0.972	0.950	0.961	0.969	0.955	0.962	0.969	0.955	0.962
Drug−route	0.943	0.931	0.937	0.954	0.933	0.943	0.936	0.939	0.937	0.936	0.939	0.937
Drug−dosage	0.921	0.928	0.924	0.933	0.931	0.932	0.925	0.950	0.937	0.925	0.950	0.937
Drug−duration	0.814	0.718	0.763	0.880	0.723	0.794	0.823	0.739	0.779	0.823	0.739	0.779
Drug−ADE[a]	0.590	0.322	0.417	0.592	0.307	0.404	0.590	0.377	0.460	0.544	0.446	0.490
Drug−reason	0.682	0.526	0.594	0.676	0.546	0.604	0.680	0.593	0.633	0.673	0.628	0.650
Overall (micro)	0.912	0.859	0.885	0.920	0.862	0.890	0.912	0.877	0.894	0.906	0.884	0.895

[a]ADE: adverse drug event.

Discussion

Principal Findings

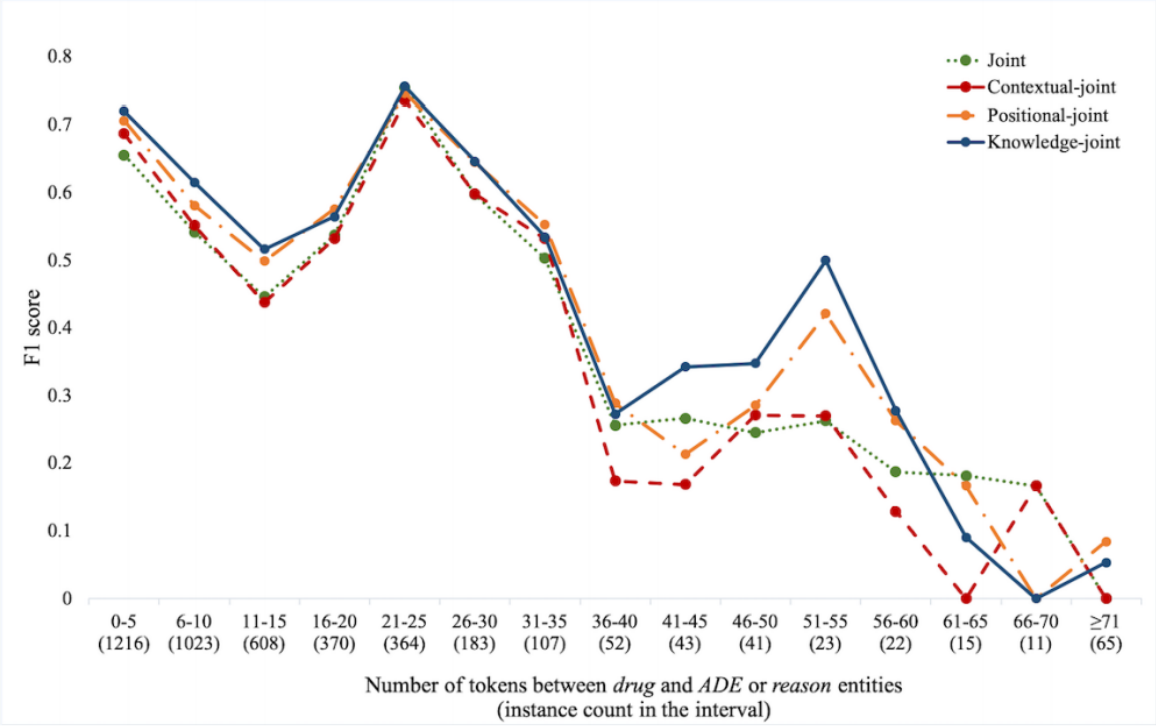
Contextualized representations (*contextual-joint*) are effective in differentiating between words and abbreviations that could have multiple meanings. For example, *ensure* and *contrast* can be understood as either a *drug* ("Ensure: 1 can PO three times daily" and "contrast-induced nephropathy") or a verb, and terms such as *blood* could either refer to a drug ("transfused 1 unit of blood"), that is, substance given to a patient, a test for the drug ("blood alcohol concentration"), or a natural occurring substance in the body ("blood pressure"). Additionally, abbreviations such as *PE* (physical examination versus pulmonary embolism) and *pcp* (primary care physician versus pneumocystis pneumonia) can have multiple expansions. In all the examples above, the *contextual-joint* correctly identifies these entities.

One prevailing challenge in ADE extraction is the presence of long-distance or intersentential relations. As shown in Table 2,

a significant portion of *drug−{ADE, reason}* in the data set is intersentential (23% of *drug−ADE* and 31.7% of *drug−reason*). These relations typically span long distances, making them more difficult to capture. To study the effectiveness of the proposed approaches over long-distance relations, we calculated the F_1 scores on *drug−{ADE, reason}* with an increasing number of tokens between entities. As shown in Figure 5, we find that the positional-*joint* model performs significantly better than the *contextual-joint* model with increasing distance between entities, suggesting that the *positional-joint* can effectively model long-distance relations.

Incorporating knowledge embeddings learned on the FAERS improved *drug−{ADE, reason}* relation extraction, especially in the case of long-distance relations or when contextual clues are insufficient. As shown in Figure 5, *the knowledge-joint* model further improved on the *positional-joint* model at all distances. The *knowledge-joint* model was also useful in cases of insufficient or ambiguous context in extracting the correct relation. For example, in the phrase "Wellbutrin - nausea and vomiting," the relation is indicated only by an uninformative hyphen, with no contextual clues to indicate the type of relation. Similarly, in "Patient had history of depression and was on elavil previously," it is unclear whether the *history of depression* was previously treated by *drug−reason* or caused by *drug−ADE* of the drug *elavil*. Furthermore, the *knowledge-joint* also helped to extract correct relations when multiple drugs and candidate *ADEs* and *reasons* are discussed in a given context. For example, in "Upon arrival, she was hypertensive and had a fever. She was given Tylenol," based on sentence construction, 2 candidate *reasons* (*hypertensive* and *fever*) may be associated with the *drugTylenol*. Knowledge is required to infer that of the two, only *fever* is related to *Tylenol*.

Figure 5. F1 scores of approaches with increasing distance between entities for relation extraction. ADE: adverse drug event.



Error Analysis

We investigated the most common error categories by entity and relation type and present these in Table 5. Most of the errors in recognizing *drug*s were due to abbreviations, misspellings, generic terms, or linguistic shorthand. For *strength* and *dosage*, these entities were often mislabeled as each other—both are often numeric quantities and used in similar contexts. For *duration* and *frequency*, most of the errors resulted from these entities being expressed in colloquial language.

Intersentential relations remain a major category of false-negative errors for all relations despite improvements from the position-attention mechanism. For *drug−{attributes}*, these errors were likely due to an insufficient number of such examples in the training data (approximately 4%). In addition to errors from intersentential relations, other important categories for false-negative *drug−{ADE, reason}* include (1) *ADE* or *reasons* expressed in generic terms, (2) *reasons* such as procedures and activities (eg, *angioplasty/stenting*) that occur infrequently in the training set, and (3) *ADE* or *reasons* expressed as abbreviations that are nonstandard or ambiguous.

False-positive errors in *drug−{ADE, reason}* mainly fall into 2 categories. In the first, one of the entities participating in the relation is negated, hypothetical, or conditional, such as when a drug is withheld to avoid an anticipated ADE (eg, contraindications). In the second, the same concept (*drug*, *ADE*, or *reason*) is mentioned multiple times in the same context, and the system associated the relation to one mention whereas the ground truth to the other. To add further complexity, these mentions may be synonyms, for example, "the pain medications (morphine, vicodin, codeine) worsened your mental status and made you delirious." With multiple possible *drug−ADE* relations, some combinations were not captured in the ground truth, resulting in false-positives that may not be true errors.

Table 5. Error analysis on our best-performing model (knowledge-joint).

Entity/relation, Error category	Text[a]	Explanation
Drug		
Abbreviation	Hyponatremia due to *HCTZ*[b]	HCTZ—abbreviated drug
Misspelled words	30 units of Lantus in addition to *humalong*	Humalog is incorrectly written as humalong
Short forms	She was given *vanco*	Vancomycin is expressed in shorthand
Generic phrase	He was advised to not take any of his *blood pressure medications*	Antihypertensives are expressed through generic terms
Strength		
Contextual ambiguity	Patient received *1 unit* of *blood*	*Strength* (*1 unit*) wrongly predicted as *dosage*; usually, the *unit* token is associated with *dosage*
Duration		
Colloquial language	Only take *Hydroxyzine as long as your rash is itching*	*Duration* is expressed colloquially
Drug − strength		
Intersentential	Continued *Carvedilol*. INR[c] initially slightly supratherapeutic, but then his home regimen of *4mg* alternating with *2mg* daily was started	Intersentential relation between carvedilol and 4 mg
Drug − ADE[d] ; Drug − reason		
Intersentential	He underwent *coronary artery bypass* x5, please see operative report for further details. He was transferred to the CSRU[e] on *Neo* with IABP[f]	Intersentential relation between neo and coronary artery bypass graft
Generic terms	Start a baby *aspirin* every day to *protect the heart*	*Reason* is expressed in generic terms
Abbreviation	*Detrol* was discontinued on suspicion that it might contribute to *AMS*	*AMS* has multiple possible expansions
Procedure	*Angioplasty* of the left tibial artery; had been on *Plavix* prior to NSTEMI[g]	Procedure angioplasty is annotated as *reason*
Contraindication	Avoiding *NSAIDs*[h] to prevent *gastrointestinal bleed*	*Drug* was not given to this patient
Negated	*Heparin*-induced *thrombocytopenia* negative	*ADE* thrombocytopenia is negated

[a]Italics indicate text that contributes to the specified error category.
[b]HCTZ: hydrochlorothiazide.
[c]INR: international normalized ratio.
[d]ADE: adverse drug event.
[e]CSRU: cardiac surgery recovery unit.
[f]IABP: intra-aortic balloon pump.
[g]NSTEMI: non–ST-elevation myocardial infarction.
[h]NSAIDs: nonsteroidal anti-inflammatory drugs.

Document-Level Analysis

From an end user perspective, the core information needed for patient care purposes is a patient-level summary of these relations, which is a unique set of extracted relations after normalization. To evaluate our system for this purpose, we measured *drug−ADE* and *drug−reason* F_1 scores by considering unique pairs of relation mentions at the document level, presented in Table 6. We observed scores at the document level to be 1 to 2 percentage points higher than the instance level.

Table 6. Document-level analysis for drug−reason and drug−adverse drug event relations.

Model	Drug−reason						Drug−ADE[a]					
	Instance level			Document level			Instance level			Document level		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Joint	0.682	0.526	0.594	0.691	0.542	0.607	0.590	0.322	0.417	0.631	0.322	0.426
Contextual-joint	0.675	0.546	0.604	0.685	0.560	0.616	0.592	0.307	0.404	0.630	0.308	0.414
Position-joint	0.680	0.593	0.633	0.692	0.611	0.649	0.590	0.376	0.460	0.647	0.384	0.482
Knowledge-joint	0.673	0.628	0.650	0.687	0.647	0.666	0.544	0.446	0.490	0.579	0.444	0.503

[a]ADE: adverse drug event.

Comparison With Previous Work

For NER, the state-of-the-art system [38] used an ensemble (committee) of 3 different methods: CRF, BiLSTM-CRF, and joint approach. They showed that the BiLSTM-CRF is the best among the single models. Thus, we compare our best model (*knowledge-joint*) with their best-performing single model and committee approach, as shown in Table 7. Overall, *the knowledge-joint* model outperformed the single model by 0.2 percentage points and achieved similar micro-F_1 to the committee approach. Notably, *the knowledge-joint* model significantly outperformed the committee approach in recognizing the crucial *ADE* (0.5 points) and *reason* (5.2 points) entities.

For relation extraction, the state-of-the-art system used the committee approach for NER, convolutional neural network – recurrent neural network (CNN-RNN) for relation classification, and postprocessing rules. Although postprocessing rules are commonly used in competitions, they often do not generalize across data sets and therefore are of limited interest in this research. As shown in Table 7, *the knowledge-joint model* outperformed the state-of-the-art approach, both with (0.4 points) and without rules (1.6 points). Notably, *the knowledge-joint* model achieved the best results and outperformed the state-of-the-art in recognizing the most crucial and difficult to extract relations: *drug−reason* (7.1 points) and *drug−ADE* (1.4 points).

Table 7. The lenient F1 scores for named entity recognition of single and state-of-the-art ensemble models compared with our best model. The lenient F1 scores for relation extraction of state-of-the-art ensemble models with and without rules, compared with our best model.

NER[a]				Relation extraction			
Entity type	BiLSTM-CRF[b] [38]	Committee [38]	Knowledge-joint	Relation type	Committee + CNN-RNN[c] [38]	Committee + CNN-RNN + Rules [38]	Knowledge-joint
Drug	0.955	0.956	0.960	N/A[d]	N/A	N/A	N/A
Strength	0.982	0.983	0.980	Drug−strength	0.964	0.972	0.975
Form	0.958	0.958	0.958	Drug−form	0.940	0.952	0.954
Frequency	0.974	0.975	0.971	Drug−frequency	0.941	0.958	0.962
Route	0.956	0.956	0.949	Drug−route	0.930	0.942	0.937
Dosage	0.943	0.948	0.946	Drug−dosage	0.923	0.935	0.937
Duration	0.856	0.862	0.846	Drug−duration	0.740	0.786	0.779
ADE[e]	0.422	0.530	0.535	Drug−ADE	0.475	0.476	0.490
Reason	0.680	0.675	0.727	Drug−reason	0.572	0.579	0.650
Overall (micro)	0.933	0.935	0.935	Overall (micro)	0.879	0.891	0.895

[a]NER: named entity recognition.

[b]BiLSTM-CRF: bidirectional long short-term memory–conditional random field.

[c]CNN-RNN: convolutional neural network–recurrent neural network.

[d]Not applicable.

[e]ADE: adverse drug event.

Limitations and Future Work

We acknowledge several limitations of this study. First, these results are specific to the n2c2 data set, which contains only intensive care unit (ICU) discharge summaries from a single health care organization. Ground truth generation and evaluation on a more diverse data set is needed to better understand the effectiveness of these proposed approaches. Second, we observed some annotation errors in the ground truth, likely due to the complex nature of the task. Further investigation is needed to quantify the prevalence of such errors and their impact on the results.

Despite achieving state-of-the-art results, the proposed system still has room for improvement, specifically in recognizing intersentential *drug−{ADE, reason}* relations. To further improve ADE extraction, we plan to explore the following research areas:

1. Although we incorporated knowledge graph embeddings, other advanced methods that use higher-order proximity and role-preserving network embedding techniques have shown promising results in the general domain. We plan to explore methods such as Edge Label Aware Network Embedding [84] rather than training separate graph embeddings for *drug−{ADE, reason}* relations.
2. The field of contextual embeddings has evolved quickly along with the release of newer language representation models trained on clinical text. We plan to explore BERT [78,85], which utilizes a transformer network to pretrain a language model for extracting better contextual word embeddings.
3. To address some of the findings from the error analysis, we plan to leverage our clinical abbreviation expansion components [86] to help resolve ambiguous mentions and also incorporate assertion recognition [26] to capture the belief state of the physician on a concept (negated, hypothetical, conditional).
4. As mentioned earlier, the proposed models performed poorly on intersentential relation extraction. To address this, we plan to explore N-ary relation extraction for cross-sentence relation extraction using graph long short-term memory networks [87].

Conclusions

We presented a system for extracting drug-centric concepts and relations that outperformed current state-of-the-art results. Experimental results showed that contextualized embeddings, position-attention mechanisms, and knowledge embeddings effectively improve deep learning-based concepts and relation extraction. Specifically, we showed the effectiveness of a position-attention mechanism in extracting long-distance relations and knowledge embeddings from the FAERS in recognizing relations where contextual clues are insufficient.

Acknowledgments

The authors wish to thank Dr Kenneth J Barker for his assistance in providing valuable feedback on the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sentence segmentation and tokenization.
[PDF File (Adobe PDF File), 392 KB - medinform_v8i7e18417_app1.pdf]

Multimedia Appendix 2
Embeddings from Language Models contextualized embeddings and large-scale information network embedding graph embeddings.
[PDF File (Adobe PDF File), 142 KB - medinform_v8i7e18417_app2.pdf]

Multimedia Appendix 3
Detailed explanation of bidirectional long short-term memory–conditional random fields and input embeddings.
[PDF File (Adobe PDF File), 369 KB - medinform_v8i7e18417_app3.pdf]

Multimedia Appendix 4
Experimental settings used in the proposed system.
[PDF File (Adobe PDF File), 106 KB - medinform_v8i7e18417_app4.pdf]

References

1. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. J Med Internet Res 2005 Mar 14;7(1):e3 [FREE Full text] [doi: 10.2196/jmir.7.1.e3] [Medline: 15829475]

2. Rosenbloom S, Stead W, Denny J, Giuse D, Lorenzi N, Brown S, et al. Generating clinical notes for electronic health record systems. Appl Clin Inform 2010 Jan 1;1(3):232-243 [FREE Full text] [doi: 10.4338/ACI-2010-03-RA-0019] [Medline: 21031148]
3. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE prevention study group. J Am Med Assoc 1995 Jul 5;274(1):29-34. [Medline: 7791255]
4. Johnson JA, Bootman JL. Drug-related morbidity and mortality. A cost-of-illness model. Arch Intern Med 1995 Oct 9;155(18):1949-1956. [Medline: 7575048]
5. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. J Am Med Assoc 1997;277(4):301-306. [Medline: 9002492]
6. Chiatti C, Bustacchini S, Furneri G, Mantovani L, Cristiani M, Misuraca C, et al. The economic burden of inappropriate drug prescribing, lack of adherence and compliance, adverse drug events in older people: a systematic review. Drug Saf 2012 Jan;35(Suppl 1):73-87. [doi: 10.1007/BF03319105] [Medline: 23446788]
7. Ahmad SR. Adverse drug event monitoring at the Food and Drug Administration. J Gen Intern Med 2003 Jan;18(1):57-60 [FREE Full text] [doi: 10.1046/j.1525-1497.2003.20130.x] [Medline: 12534765]
8. Hazell L, Shakir SA. Under-reporting of adverse drug reactions: a systematic review. Drug Saf 2006;29(5):385-396. [doi: 10.2165/00002018-200629050-00003] [Medline: 16689555]
9. Tatonetti NP, Fernald GH, Altman RB. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. J Am Med Inform Assoc 2012;19(1):79-85 [FREE Full text] [doi: 10.1136/amiajnl-2011-000214] [Medline: 21676938]
10. Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. Br Med J 2004 Jul 3;329(7456):15-19 [FREE Full text] [doi: 10.1136/bmj.329.7456.15] [Medline: 15231615]
11. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. Clin Pharmacol Ther 2012 Aug;92(2):228-234 [FREE Full text] [doi: 10.1038/clpt.2012.54] [Medline: 22713699]
12. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation analysis for testing drug safety signals using unstructured clinical notes. J Biomed Semantics 2012 Apr 24;3(Suppl 1):S5 [FREE Full text] [doi: 10.1186/2041-1480-3-S1-S5] [Medline: 22541596]
13. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clin Pharmacol Ther 2013 Jun;93(6):547-555 [FREE Full text] [doi: 10.1038/clpt.2013.47] [Medline: 23571773]
14. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data 2016 May 10;3:160026 [FREE Full text] [doi: 10.1038/sdata.2016.26] [Medline: 27193236]
15. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 2012 Jun;91(6):1010-1021 [FREE Full text] [doi: 10.1038/clpt.2012.50] [Medline: 22549283]
16. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. Ann Intern Med 2010 Nov 2;153(9):600-606. [doi: 10.7326/0003-4819-153-9-201011020-00010] [Medline: 21041580]
17. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: a review. J Biomed Inform 2015 Apr;54:202-212 [FREE Full text] [doi: 10.1016/j.jbi.2015.02.004] [Medline: 25720841]
18. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-681 [FREE Full text] [doi: 10.1093/jamia/ocu041] [Medline: 25755127]
19. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. Drug Saf 2014 May;37(5):343-350 [FREE Full text] [doi: 10.1007/s40264-014-0155-x] [Medline: 24777653]
20. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug Saf 2014 Oct;37(10):777-790 [FREE Full text] [doi: 10.1007/s40264-014-0218-z] [Medline: 25151493]
21. Odgers D, Harpaz R, Callahan A, Stiglic G, Shah N. Analyzing search behavior of healthcare professionals for drug safety surveillance. Biocomputing 2014(2014):306-317 [FREE Full text] [doi: 10.1142/9789814644730_0030]
22. White RW, Harpaz R, Shah NH, DuMouchel W, Horvitz E. Toward enhanced pharmacovigilance using patient-generated data on the internet. Clin Pharmacol Ther 2014 Aug;96(2):239-246 [FREE Full text] [doi: 10.1038/clpt.2014.77] [Medline: 24713590]
23. Abernethy DR, Woodcock J, Lesko LJ. Pharmacological mechanism-based drug safety assessment and prediction. Clin Pharmacol Ther 2011 Jun;89(6):793-797. [doi: 10.1038/clpt.2011.55] [Medline: 21490594]

24. Chiang A, Butte A. Data-driven methods to discover molecular determinants of serious adverse drug events. Clin Pharmacol Ther 2009 Mar;85(3):259-268 [FREE Full text] [doi: 10.1038/clpt.2008.274] [Medline: 19177064]
25. Vilar S, Harpaz R, Chase H, Costanzi S, Rabadan R, Friedman C. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. J Am Med Inform Assoc 2011 Dec;18(Suppl 1):i73-i80 [FREE Full text] [doi: 10.1136/amiajnl-2011-000417] [Medline: 21946238]
26. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
27. Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. Semantic Scholar. 2017. URL: https://pdfs.semanticscholar.org/5b8a/7b11b987ddeb865dbf3aaa7b745a86ea5bf0.pdf [accessed 2020-06-22]
28. Li J, Sun Y, Johnson R, Sciaky D, Wei C, Leaman R, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw068] [Medline: 27161011]
29. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019 Jan;42(1):99-111 [FREE Full text] [doi: 10.1007/s40264-018-0762-z] [Medline: 30649735]
30. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 1;27(1):3-12. [doi: 10.1093/jamia/ocz166]
31. Xu J, Wu Y, Zhang Y, Wang J, Lee HJ, Xu J. CD-REST: a system for extracting chemical-induced disease relation in literature. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw036] [Medline: 27016700]
32. Finkel J, Dingare S, Manning CD, Nissim M, Alex B, Grover C. Exploring the boundaries: gene and protein identification in biomedical text. BMC Bioinformatics 2005;6(Suppl 1):S5 [FREE Full text] [doi: 10.1186/1471-2105-6-S1-S5] [Medline: 15960839]
33. Wei C, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw032] [Medline: 26994911]
34. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform 2012 Oct;45(5):885-892 [FREE Full text] [doi: 10.1016/j.jbi.2012.04.008] [Medline: 22554702]
35. Sutton C. An introduction to conditional random fields. FNT in Mach Learn 2012;4(4):267-373 [FREE Full text] [doi: 10.1561/2200000013]
36. Andrew AM. An introduction to support vector machines and other kernel‐based learning methods. Kybernetes 2001 Feb;30(1):103-115. [doi: 10.1108/k.2001.30.1.103.6]
37. Dandala B, Joopudi V, Devarakonda M. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. Drug Saf 2019 Jan;42(1):135-146. [doi: 10.1007/s40264-018-0764-x] [Medline: 30649738]
38. Wei Q, Ji Z, Li Z, Du J, Wang J, Xu J, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc 2020 Jan 1;27(1):13-21 [FREE Full text] [doi: 10.1093/jamia/ocz063] [Medline: 31135882]
39. Li F, Liu W, Yu H. Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning. JMIR Med Inform 2018 Nov 26;6(4):e12159 [FREE Full text] [doi: 10.2196/12159] [Medline: 30478023]
40. Ju M, Nguyen N, Miwa M, Ananiadou S. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. J Am Med Inform Assoc 2020 Jan 1;27(1):22-30 [FREE Full text] [doi: 10.1093/jamia/ocz075] [Medline: 31197355]
41. Dai H, Su C, Wu C. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. J Am Med Inform Assoc 2020 Jan 1;27(1):47-55. [doi: 10.1093/jamia/ocz120] [Medline: 31334805]
42. Wunnava S, Qin X, Kakar T, Sen C, Rundensteiner EA, Kong X. Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. Drug Saf 2019 Jan;42(1):113-122. [doi: 10.1007/s40264-018-0765-9] [Medline: 30649736]
43. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting adverse drug events with rapidly trained classification models. Drug Saf 2019 Jan;42(1):147-156 [FREE Full text] [doi: 10.1007/s40264-018-0763-y] [Medline: 30649737]
44. Yang X, Bian J, Fang R, Bjarnadottir R, Hogan W, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc 2020 Jan 1;27(1):65-72. [doi: 10.1093/jamia/ocz144] [Medline: 31504605]
45. Chalapathy R, Borzeshi E, Piccardi M. Bidirectional LSTM-CRF for Clinical Concept Extraction. arXiv 2016 epub ahead of print(1611.08373) [FREE Full text]

46. Kim Y, Jernite Y, Sontag D, Rush A. Character-Aware Neural Language Models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016 Presented at: AAAI'16; February 12-17, 2016; Phoenix, Arizona, USA URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewFile/12489/12017
47. Li F, Zhang M, Fu G, Ji D. A neural joint model for entity and relation extraction from biomedical text. BMC Bioinformatics 2017 Mar 31;18(1):198 [FREE Full text] [doi: 10.1186/s12859-017-1609-9] [Medline: 28359255]
48. Caruana R. Multitask learning. Mach Learn 1997;28(1):41-75. [doi: 10.1007/978-1-4615-5529-2_5]
49. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. In: Proceedings of the 3rd International Conference for Learning Representations. 2015 Presented at: ICLR'15; May 7-9, 2015; San Diego, CA, USA.
50. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. Adv Neural Inf Process Syst 2017:5998-6008 [FREE Full text]
51. Wang W, Yang N, Wei F, Chang B, Zhou M. Gated Self-Matching Networks for Reading Comprehension and Question Answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017 Presented at: ACL'17; July 30-August 4, 2017; Vancouver, Canada. [doi: 10.18653/v1/p17-1018]
52. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, et al. Attention-based bidirectional long short-term memory networks for relation classification. 54th Annu Meet Assoc Comput Linguist ACL 2016;2:207-212 [FREE Full text] [doi: 10.18653/v1/p16-2034]
53. Zhang Y, Zhong V, Chen D, Angeli G, Manning C. Position-Aware Attention and Supervised Data Improve Slot Filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: EMNLP'17; September 7-11, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/d17-1004]
54. Dai D, Xiao X, Lyu Y, Dou S, She Q, Wang H. Joint Extraction of Entities and Overlapping Relations Using Position-Attentive Sequence Labeling. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Presented at: Proc AAAI Conf Artif Intell ;33; January 27-February 1, 2019; Honolulu, Hawaii, USA. [doi: 10.1609/aaai.v33i01.33016300]
55. Christopoulou F, Tran T, Sahu S, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. J Am Med Inform Assoc 2020 Jan 1;27(1):39-46 [FREE Full text] [doi: 10.1093/jamia/ocz101] [Medline: 31390003]
56. Zhou H, Lang C, Liu Z, Ning S, Lin Y, Du L. Knowledge-guided convolutional networks for chemical-disease relation extraction. BMC Bioinformatics 2019 May 21;20(1):260 [FREE Full text] [doi: 10.1186/s12859-019-2873-7] [Medline: 31113357]
57. Ding R, Xie P, Zhang X, Lu W, Li L, Si L. A Neural Multi-digraph Model for Chinese NER with Gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL'19; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/P19-1141]
58. Shen Y, Deng Y, Yang M, Li Y, Du N, Fan W, et al. Knowledge-Aware Attentive Neural Network for Ranking Question Answer Pairs. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018 Presented at: SIGIR'18; July 8-12, 2018; Ann Arbor, Michigan. [doi: 10.1145/3209978.3210081]
59. Li P, Mao K, Yang X, Li Q. Improving Relation Extraction with Knowledge-Attention. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019 Presented at: EMNLP-IJCNLP'19; November 3-7, 2019; Hong Kong, China. [doi: 10.18653/v1/D19-1022]
60. Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. Translating Embeddings for Modeling Multi-Relational Data. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013 Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, USA URL: https://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data [doi: 10.5555/2999792.2999923]
61. Wang Z, Zhang J, Feng J, Chen Z. Knowledge Graph Embedding by Translating on Hyperplanes. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014 Presented at: AAAI'14; July 27-31, 2014; Québec City, Québec, Canada URL: https://persagen.com/files/misc/wang2014knowledge.pdf
62. Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015 Presented at: AAAI'15; January 25-30, 2015; Austin, Texas, USA URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewFile/9571/9523
63. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online Learning of Social Representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014 Presented at: KDD'14; August 24-27, 2014; New York, USA. [doi: 10.1145/2623330.2623732]
64. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'18; August 13-17, 2016; San Francisco, USA. [doi: 10.1145/2939672.2939754]
65. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. LINE: Large-Scale Information Network Embedding. In: Proceedings of the 24th International Conference on World Wide Web. 2015 Presented at: WWW'15; May 18-22, 2015; Florence, Italy. [doi: 10.1145/2736277.2741093]

66. Gao M, Chen L, He X, Zhou A. BiNE: Bipartite Network Embedding. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018 Presented at: SIGIR'18; July 8-12, 2018; Ann Arbor, Michigan. [doi: 10.1145/3209978.3209987]
67. Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, et al. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. J Am Med Inform Assoc 2020 Jan 1;27(1):56-64. [doi: 10.1093/jamia/ocz141] [Medline: 31591641]
68. Loper E, Bird S. NLTK: The Natural Language Toolkit. 2002. URL: https://www.nltk.org/ [accessed 2020-06-22]
69. Honnibal M, Montani I. spaCy v2. GitHub. 2017. URL: https://github.com/explosion/spaCy/issues/1555 [accessed 2020-06-22]
70. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford coreNLP natural language processing toolkit. Assoc Comput Linguist Syst Demonstr 2014:60. [doi: 10.3115/v1/p14-5010]
71. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform 2015 Oct;57:28-37 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.010] [Medline: 26187250]
72. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. NIPS Proceedings. 2013. URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf [accessed 2020-06-22]
73. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014 Presented at: EMNLP'14; October 25-29, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
74. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017 Presented at: EACL'17; April 3-7, 2017; Valencia, Spain. [doi: 10.18653/v1/e17-2068]
75. Dai X, Karimi S, Hachey B, Paris C. Using Similarity Measures to Select Pretraining Data for NER. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL'19; July 5-10, 2019; Minneapolis, Minnesota p. 1460-1470 URL: https://www.aclweb.org/anthology/N19-1149/ [doi: 10.18653/v1/N19-1149]
76. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
77. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: MAACL'18; June 1-6, 2018; New Orleans, Louisiana. [doi: 10.18653/v1/n18-1202]
78. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL'19; June 2-7, 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/N19-1423]
79. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019 Nov 1;26(11):1297-1304. [doi: 10.1093/jamia/ocz096] [Medline: 31265066]
80. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
81. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv 2015 epub ahead of print(1508.01991) [FREE Full text]
82. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation Classification via Convolutional Deep Neural Network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014 Presented at: COLING'14; August 23-29, 2014; Dublin, Ireland URL: https://www.aclweb.org/anthology/C14-1220/
83. Edgington ES. Approximate randomization tests. J Psychol Interdiscip Appl 1969 Jul;72(2):143-149. [doi: 10.1080/00223980.1969.10543491]
84. Goyal P, Hosseinmardi H, Ferrara E, Galstyan A. Capturing edge attributes via network embedding. IEEE Trans Comput Soc Syst 2018 Dec;5(4):907-917. [doi: 10.1109/tcss.2018.2877083]
85. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
86. Joopudi V, Dandala B, Devarakonda M. A convolutional route to abbreviation disambiguation in clinical text. J Biomed Inform 2018 Oct;86:71-78 [FREE Full text] [doi: 10.1016/j.jbi.2018.07.025] [Medline: 30118854]
87. Peng N, Poon H, Quirk C, Toutanova K, Yih W. Cross-sentence n-ary relation extraction with graph LSTMs. Transact Assoc Comput Ling 2017;-:101-115. [doi: 10.1162/tacl_a_00049]

Abbreviations

ADE: adverse drug event
AEOLUS: adverse event open learning through universal standardization
BERT: Bidirectional Enconder Representations from Transformers
BiLSTM-CRF: bidirectional, long short-term memory–conditional random fields
BIO: beginning, inside, and outside
CNN: convolutional neural network
CRF: conditional random field
CUI: concept unique identifier
EHR: electronic health record
ELMo: Embeddings from Language Models
FAERS: Food and Drug Administration Adverse Event Reporting System
LINE: large-scale information network embedding
MADE 1.0: Medication and Adverse Drug Events from Electronic Health Records
MIMIC-III: Medical Information Mart for Intensive Care III
n2c2: 2018 National NLP Clinical Challenges
NER: named entity recognition
NLP: natural language processing
SBD: sentence boundary detection
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 25.02.20; peer-reviewed by T Muto, S Doan; comments to author 28.04.20; revised version received 12.05.20; accepted 13.05.20; published 10.07.20.

Please cite as:
Dandala B, Joopudi V, Tsou CH, Liang JJ, Suryanarayanan P
Extraction of Information Related to Drug Safety Surveillance From Electronic Health Record Notes: Joint Modeling of Entities and Relations Using Knowledge-Aware Neural Attentive Models
JMIR Med Inform 2020;8(7):e18417
URL: https://medinform.jmir.org/2020/7/e18417
doi:10.2196/18417
PMID:32459650

Original Paper

Predicting Current Glycated Hemoglobin Levels in Adults From Electronic Health Records: Validation of Multiple Logistic Regression Algorithm

Zakhriya Alhassan[1,2], MSc; David Budgen[1], PhD; Riyad Alshammari[3,4], PhD; Noura Al Moubayed[1], PhD

[1]Department of Computer Science, Durham University, Durham, United Kingdom

[2]Computer Science Department, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

[3]College of Public Health and Health Informatics, Health Informatics Department, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

[4]King Abdullah International Medical Research Center, Ministry of the National Guard - Health Affairs, Riyadh, Saudi Arabia

Corresponding Author:
Zakhriya Alhassan, MSc
Department of Computer Science
Durham University
Mountjoy Centre
Stockton Road
Durham, DH1 3LE
United Kingdom
Phone: 44 191 3341724
Email: zakhriya.n.alhassan@durham.ac.uk

Abstract

Background: Electronic health record (EHR) systems generate large datasets that can significantly enrich the development of medical predictive models. Several attempts have been made to investigate the effect of glycated hemoglobin (HbA_{1c}) elevation on the prediction of diabetes onset. However, there is still a need for validation of these models using EHR data collected from different populations.

Objective: The aim of this study is to perform a replication study to validate, evaluate, and identify the strengths and weaknesses of replicating a predictive model that employed multiple logistic regression with EHR data to forecast the levels of HbA_{1c}. The original study used data from a population in the United States and this differentiated replication used a population in Saudi Arabia.

Methods: A total of 3 models were developed and compared with the model created in the original study. The models were trained and tested using a larger dataset from Saudi Arabia with 36,378 records. The 10-fold cross-validation approach was used for measuring the performance of the models.

Results: Applying the method employed in the original study achieved an accuracy of 74% to 75% when using the dataset collected from Saudi Arabia, compared with 77% obtained from using the population from the United States. The results also show a different ranking of importance for the predictors between the original study and the replication. The order of importance for the predictors with our population, from the most to the least importance, is age, random blood sugar, estimated glomerular filtration rate, total cholesterol, non–high-density lipoprotein, and body mass index.

Conclusions: This replication study shows that direct use of the models (calculators) created using multiple logistic regression to predict the level of HbA_{1c} may not be appropriate for all populations. This study reveals that the weighting of the predictors needs to be calibrated to the population used. However, the study does confirm that replicating the original study using a different population can help with predicting the levels of HbA_{1c} by using the predictors that are routinely collected and stored in hospital EHR systems.

(JMIR Med Inform 2020;8(7):e18963) doi:10.2196/18963

KEYWORDS

glycated hemoglobin; HbA1c; prediction; electronic health records; diabetes; differentiated replication; EHR; hemoglobin; logistic regression; medical informatics

Introduction

Diabetes is a growing medical condition worldwide. Globally, the estimated number of diabetic patients in 2017 was 425 million, and it is expected to be more than 629 million by 2045, an increase of more than 48%. The number of people with borderline diabetes is also rapidly increasing. According to the International Diabetes Federation (IDF), there are 352 million people worldwide who are at risk of developing diabetes [1]. The latest estimates indicate that 35.3% of the adults in the United Kingdom and the United States have prediabetes [2].

Type 2 diabetes mellitus (T2DM) is the most common form of diabetes, accounting for 91% to 95% of all cases [3]. T2DM is difficult to diagnose in its early stages because it does not have clear clinical symptoms. As a result of the slow development of its symptoms, it often stays undetected for a long time [4]. The IDF estimates that half of people with diabetes do not know or feel that they are developing diabetes [1].

Hemoglobin is responsible for transporting oxygen throughout the body's cells and, when joined with the glucose within the blood, it forms glycated hemoglobin (HbA_{1c}) [5,6]. The International Expert Committee, with members from the American Diabetes Association (ADA), the European Association for the Study of Diabetes, and the International Diabetes Federation [7,8], recommends the use of the glycated hemoglobin test to identify adults with a high risk of diabetes [9].

An elevation of HbA_{1c} level in the blood can be related to chronic complications and lead to serious health conditions [10]. Patients with HbA_{1c} levels of 5.5% to 6.0% have a substantial risk of developing diabetes, increased by 25% compared with patients with HbA_{1c} levels less than 5.5%. Furthermore, patients with HbA_{1c} levels of more than 6.0% have a 50% chance of developing T2DM over the next 5 years. Those patients are at 20 or more times higher risk than patients who have a level of 5.0% or less [11].

A study by Huang et al [12] showed that patients with HbA_{1c} levels of 5.7% to 6.5% are likely to develop diabetes in 2.49 years. Not only that, but the trend of the HbA_{1c} test has been shown to be an important factor for predicting mortality for patients with T2DM [13]. Furthermore, nondiabetic people with an elevated HbA_{1c} level have an increased risk of cardiovascular disease [9,14]. Hence, studies suggest that patients with and without diabetes with raised levels of HbA_{1c} should be clinically checked and monitored as a preventive intervention to avoid developing T2DM or cardiovascular diseases [14,15].

Many studies have investigated the correlation between HbA_{1c} and clinical variables using statistical and mathematical approaches [16-19]. However, we are not aware of any that have performed replications of the predictive models on different populations. In this paper, we investigate building statistical models that predict the probability of patients having an elevated level of HbA_{1c}. We employ comparative statistical models similar to the models used by Wells et al [2] and apply them to a larger electronic health record (EHR) dataset collected from King Abdullah International Medical Research Center (KAIMRC) [20,21] in Saudi Arabia.

The work by Wells et al [2], which we refer to in this paper as the original study, focused on predicting the level of HbA_{1c} for patients who were not previously diagnosed with diabetes or taking diabetes medications. The data were extracted from the EHR database of Wake Forest Baptist Medical Center in the United States. The authors applied a multiple logistic regression model to create a mathematical equation for calculating the level of HbA_{1c} (≥5.7). The predictors used in the equation were chosen from a list of theoretically associated hyperglycemia variables (laboratory measurements, medication categories, diagnosis, vital signs, demographics, family history, and social history variables). After reducing the model's variables using Harrell's model approximation method [22] and removing variables that caused collinearity, the final equation associated 8 independent variables with the result of the HbA_{1c} blood test. Restricted cubic splines (RCS) with 3 knots were used for fitting the continuous predictors into the model [2]. The calculator achieved an accuracy of 77%.

The independent replication of empirical studies is widely regarded as being an essential underpinning of the scientific paradigm. Successful replication of a study by other researchers is considered to be an important step in verifying the original findings and helping to determine how widely they apply.

While the vocabulary associated with replication varies across disciplines [23], the terms employed by Lindsay and Ehrenberg [24] appear to be widely used and recognized, so they will be used in this paper. Lindsay and Ehrenberg categorize replication studies as either (1) close replications or (2) differentiated replications.

First, a close replication seeks to repeat the original study in a way that keeps all the "known conditions of the study the same or very similar" [24]. Hence, such a study employs the same forms of measurement, sampling, and analysis as the original, while also seeking to keep the profile of any set of participants as close to the original as possible. A close replication aims to test the hypothesis that, when a given study is repeated under the same experimental conditions as the original study, it should produce the same (or nearly the same) result.

Second, a differentiated replication introduces known variations into what Lindsay and Ehrenberg term "fairly major aspects of the conditions of the study" [24]. Differentiated replications provide a test of how widely the original findings can be generalized, their scope, and the conditions under which they may not hold. For a differentiated replication, therefore, it is expected that some changes in the outcomes are likely to arise,

and the question of interest is to what extent and in what form these outcome changes occur.

In an ideal situation, one or more close replications would be used to validate the findings of an original study, followed by a set of differentiated replications used to scope out the extent of their validity by varying different conditions.

For any replication study, it is possible to vary one or more factors from those factors that characterize the way that the study was performed. These may include the team performing the replication, the analysis process, the type of data employed, and the population from which the data were derived. As this study involves analyzing data collected from a human population rather than conducting an experiment or trial, we can expect that using a different team to perform a replication should have no effect. Hence, for a close replication it would be appropriate to use the same analysis tool with EHRs of the same form as used in the original study, but pertaining to a different sample of participants drawn from the same general population used in the original study.

For the differentiated replication reported here, we have used the same form of analysis, but have applied this to a set of EHRs that were derived from a different population. The differences between the forms of the EHRs constituted one difference, but these differences were relatively small. The main difference in the studies arose from the population used. As with the original study, the selection of participants was largely driven by availability. We therefore expected that it was quite possible that there would be some differences in the outcomes, and our main goal was to investigate the extent and form of those differences.

Methods

Conduct of the Replication Study

The KAIMRC dataset was collected by the Ministry of National Guard Health Affairs from the EHR systems of National Guard Hospitals in Saudi Arabia for the period from 2016 to the end of 2018. The dataset was then labelled according to the ADA guidelines. Patients with an HbA_{1c} level of 5.7% or more are considered to have an elevated HbA_{1c} and those with lower levels than that are considered normal. The predictors that were selected by the authors of the original study for calculating the level of HbA_{1c}, listed in Table 1, were employed in this study, except for race and smoking status. Taking into account that most of the data samples in the KAIMRC dataset are from the same race, the race variable can be omitted, as it has zero variance [25]. Smoking status information is absent from the KAIMRC dataset. However, in the original model used by Wells et al, this was ranked as having the lowest importance of all the predictors. The BMI and non–high-density lipoprotein measures were also absent. However, both can be calculated by using the formulae presented in Multimedia Appendix 1.

Table 1. Predictors available in the original study versus King Abdullah International Medical Research Center datasets.

Predictors	Original study dataset	KAIMRC[a] dataset
Age	√	√
Body mass index	√	√ (calculated)
Estimated glomerular filtration rate	√	√
Random blood sugar (glucose) level	√	√
Non–high density lipoprotein	√	√ (calculated)
Total cholesterol	√	√
Race	√	x
Smoking status	√	x

[a]KAIMRC: King Abdullah International Medical Research Center, Saudi Arabia.

In this study we followed the same sampling approach used in original study. For inpatient visits, only the first day's data were considered, and in cases of missing values, the first available values for the visit were used. Samples for patients with values of <1% for HbA_{1c} were simply considered to be erroneous readings and were excluded. Similar to the original study, patients diagnosed with diabetes were eliminated from the development dataset (refer to Multimedia Appendix 2 for diabetes diagnostic codes). We avoided intensive interpretation for handling the missing values. Samples with one or more completely missing values were also excluded. This resulted in decreasing the dataset size from the 262,559 samples originally collected to 36,378 samples. Figure 1 shows the detailed preprocessing tasks performed prior to building the statistical models.

The descriptive statistics for the KAIMRC experimental dataset and the dataset used by Wells et al are shown in Table 2. The units used for recording lab tests can differ according to the laboratory guidelines followed by each country. The KAIMRC dataset uses different units than the ones used in the original study for some variables. For instance, the total cholesterol level is measured in milligrams per deciliter (mg/dL) in the original study's dataset, and in millimoles per liter (mmol/L) in the dataset from the KAIMRC labs. Therefore, the descriptive statistics contain the values using both units. When developing the predictive models, the authors converted the units using the appropriate formulae (see Multimedia Appendix 3). However, the conversion task can be avoided to reduce data preprocessing complexity, as it should not affect the prediction performance for the logistic regression models.

Figure 1. Dataset preprocessing details. HbA_{1c}: glycated hemoglobin.

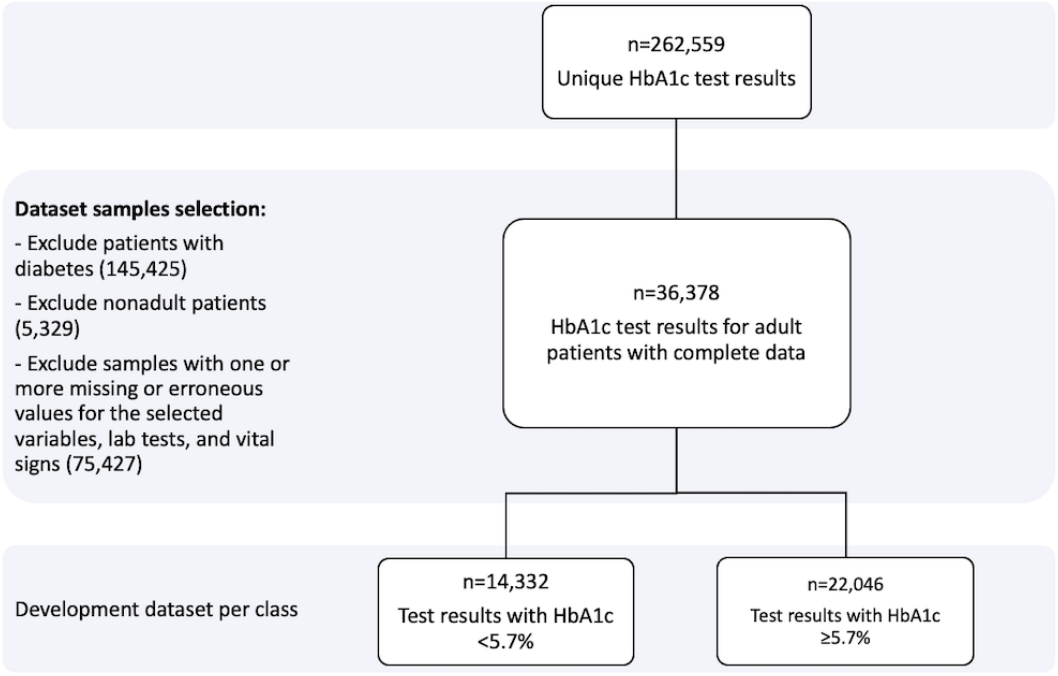


Table 2. Descriptive statistics for King Abdullah International Medical Research Center and original study datasets.

Variables[a]	KAIMRC[b] dataset			Original study[c] dataset	
	HbA_{1c}[d] <5.7% (n=14,332)	HbA_{1c} ≥5.7% (n=22,046)	*P* value	HbA_{1c} <5.7% (n=16,743)	HbA_{1c} ≥5.7% (n=5892)
Age (years), mean (SD)	45.5 (17.01)	60.5 (14.13)	<.001	48.1 (15.4)	54.8 (14.0)
BMI (kg/m^2), mean (SD)	29.61 (10.74)	31.50 (12.13)	<.001	30.1 (7.44)	33.0 (8.41)
eGFR[e] ($mL/min/1.73\ m^2$), mean (SD)	93.40 (35.19)	82.02 (28.86)	<.001	92.0 (33.0)	87.9 (30.8)
RBS[f]			<.001		
RBS (mmol/L), mean (SD)	5.47 (1.28)	8.30 (4.30)		4.9 (0.7)	5.3 (0.9)
RBS (mg/dL), mean (SD)	98.5 (23.00)	149.4 (77.47)		88.4 (12.7)	96.1 (16.0)
Cholesterol			<.001		
Cholesterol (mmol/L), mean (SD)	4.59 (1.19)	4.17 (1.16)		4.80 (1.01)	4.96 (1.11)
Cholesterol (mg/dL), mean (SD)	177.49 (46.01)	161.25 (44.85)		186 (39.4)	192 (43.1)
Non-HDL[g]			<.001		
Non-HDL (mmol/L), mean (SD)	2.85 (1.06)	2.49 (0.99)		3.49 (0.96)	3.72 (1.07)
Non-HDL (mg/dL), mean (SD)	110.2 (40.99)	96.28 (38.28)		135 (37.4)	144 (41.7)

[a]Refer to Multimedia Appendix 3 for unit conversion formulae.

[b]KAIMRC: King Abdullah International Medical Research Center, Saudi Arabia.

[c]Wake Forest Baptist Medical Center, North Carolina, United States.

[d]HbA_{1c}: glycated hemoglobin.

[e]eGFR: estimated glomerular filtration rate.

[f]RBS: random blood sugar.

[g]HDL: high-density lipoproteins.

Study Design

A complete validation of Wells et al's calculator using our dataset was not possible due to the absence of the smoking status variable. To validate the approach used in the original study, 3 predictive models (PMs) were built, trained, and tested using the KAIMRC dataset. All models employ multiple logistic regression to create the calculator by associating the chosen and available predictors. After discussion with the authors of the original study, we structured the models as PM1, PM2, and PM3.

PM1 was designed to be as close as possible to the original study's model. It uses the predictors chosen in the original study: age, BMI, random blood sugar (RBS), non–high-density lipoprotein (non-HDL), cholesterol, and estimated glomerular filtration rate (eGFR). The continuous predictors are fitted to the model using RCS with 3 knots.

PM2 was designed using the same predictors used in PM1 but without RCS fitting.

PM3 was designed after excluding the predictors with the least importance in PM1 and PM2, using a reduced number of predictors and fitted using RCS with 5 knots. The choice of the number of knots for this model was determined by using Stone's recommendation [26].

The 3 models were validated using the 10-fold cross-validation approach. The measure used to evaluate and compare the results with the original study was the concordance statistic, which is equal to area under the receiver operating characteristic (AUR ROC) curve [27]. To assist with future comparisons, we report measures commonly used for medical research, such as precision, recall, and F1, in the model evaluation. The data preparations are undertaken using Python (version 3.7; Python Software Foundation). The model building and the analysis are carried out in R (version 3.6.0; The R Foundation) using the regression modeling strategies package.

Results

The development data subset size used for training, testing, and validating the models after data preprocessing was 36,378 samples. Most medical datasets are imbalanced with a majority normal population [28], but 60.60% (22,046/36,378) of KAIMRC dataset patients were found to have elevated levels of HbA_{1c} (≥5.7%), and 39.40% (14,332/36,378) of patients had a normal HbA_{1c} level (<5.7%).

Details of the 3 models (PM1, PM2, and PM3) used for the purpose of validating and evaluating the original study are shown in Table 3. This study explores multiple logistic regression models using different numbers of variables, with and without RCS, and with different numbers of knots. PM1 (using a complete set of variables fitted using RCS) achieves an average accuracy of 73.67% and 95% CI of 74% to 77% with a well-calibrated curve. A similar model (PM2), but not fitted using RCS, shows improved accuracy, with an average accuracy of 74.04% and the same 95% CI of 74% to 77%. However, the calibration curve shows better calibration when applying RCS into the models, as shown in Figures 2 and 3.

Table 3. Performance of models for glycated hemoglobin elevation prediction.

Model	Variables used	Number of RCS[a] knots	AUR ROC[b]	95% CI	Recall	Precision	F1
PM[c]1	Complete[d]	3	73.67	74.71-77.51	85.24	77.58	81.23
PM2	Complete	N/A[e]	74.04	74.35-77.16	82.18	78.76	80.43
PM3	Reduced[f]	5	74.73	75.38-78.15	84.40	78.80	81.50

[a]RCS: restricted cubic splines.

[b]AUR ROC: area under the receiver operating characteristic.

[c]PM: predictive model.

[d]All variables (age, random blood sugar, cholesterol, non–high-density lipoproteins, estimated glomerular filtration rate, and BMI).

[e]N/A: not applicable.

[f]Reduced variables (age, random blood sugar, cholesterol, non–high-density lipoproteins, and estimated glomerular filtration rate).

Figure 2. The calibration curve for PM1. HbA_{1c}: glycated hemoglobin. PM: predictive model.

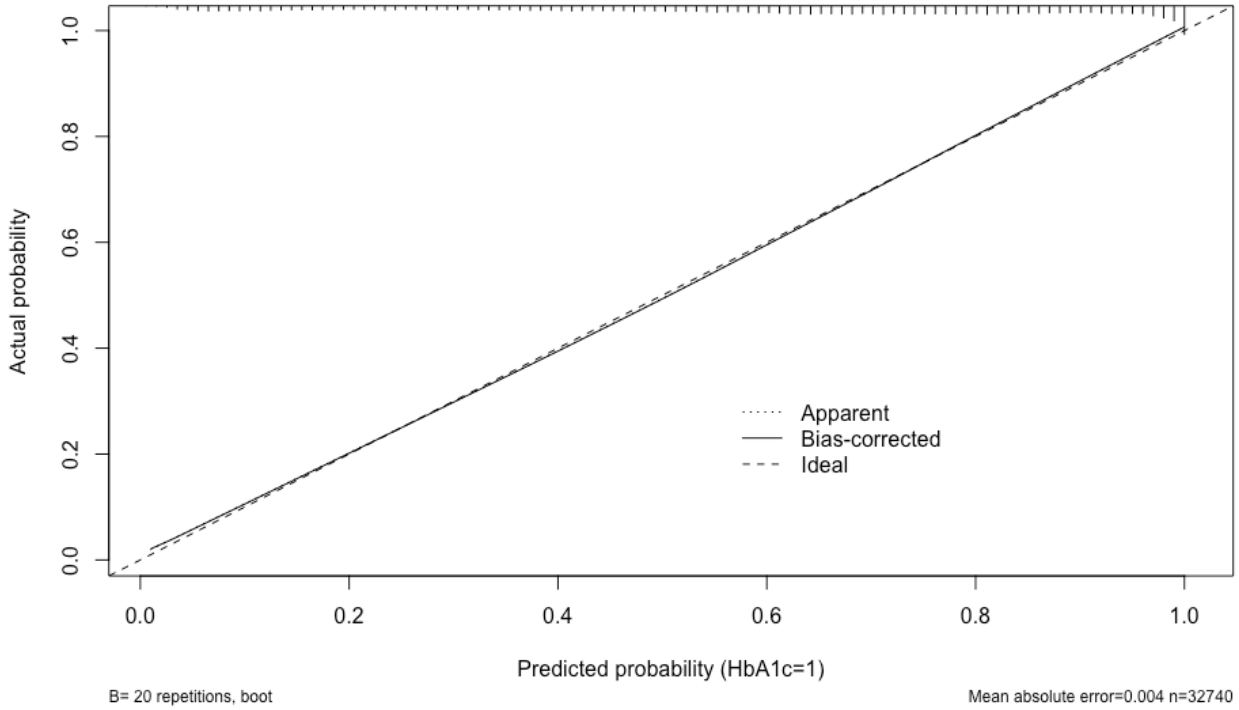


Figure 3. The calibration curve for PM2. HbA_{1c}: glycated hemoglobin. PM: predictive model.

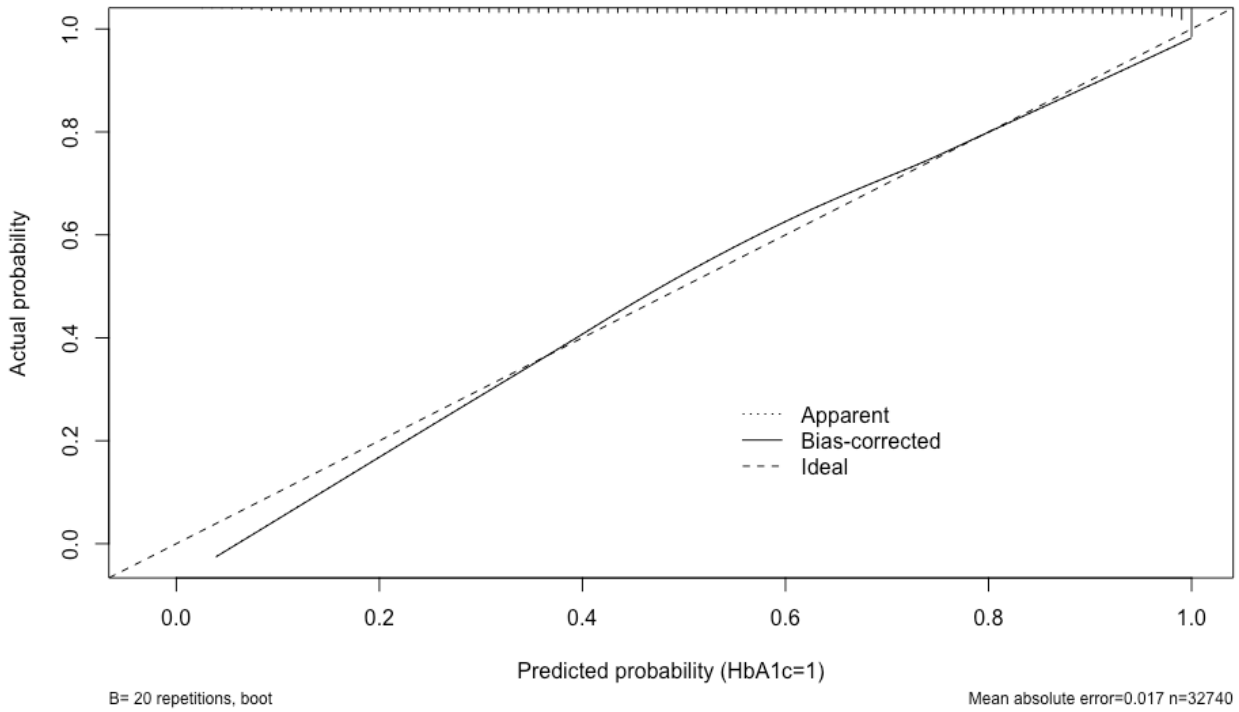
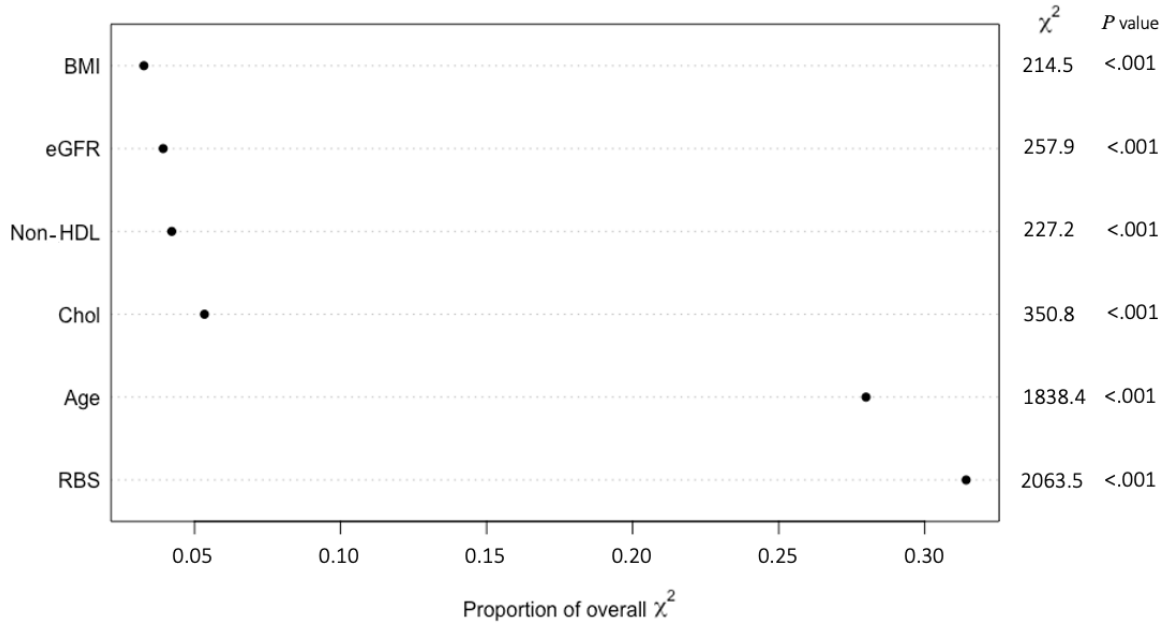


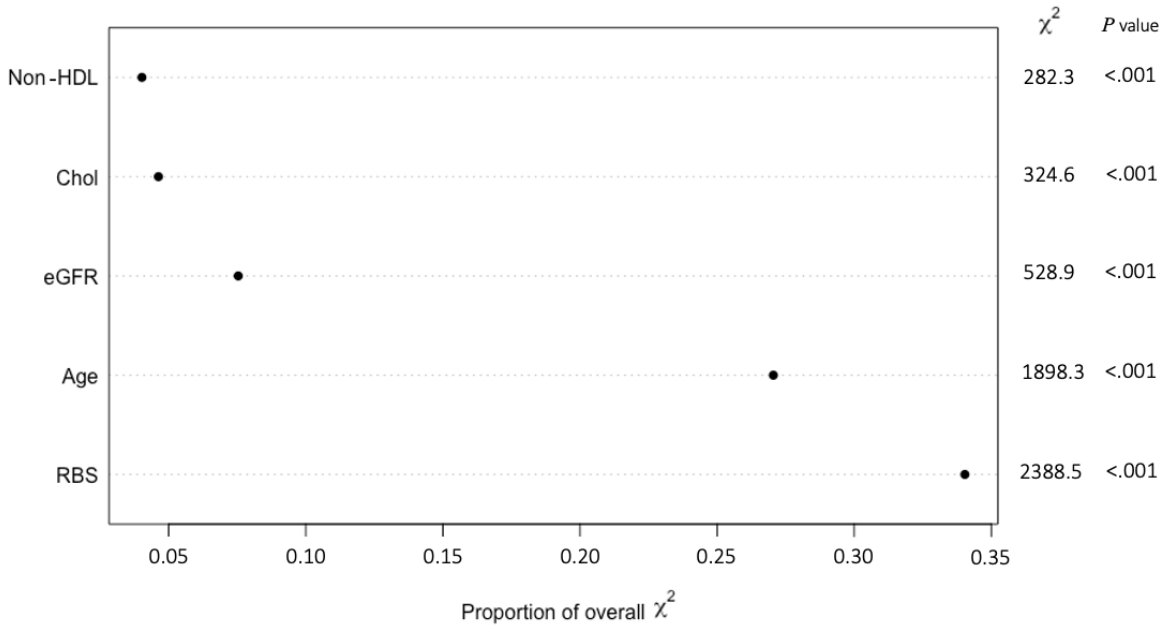
Figure 4 shows the ranking of importance for the variables used in the PM1 model. PM1 shows a different order of importance for the predictors than the order obtained from the original study. Age and RBS are of great importance in both studies. However, BMI is of the lowest importance when using the KAIMRC population, whereas in the original study it was ranked second.

Figure 4. Order of importance of predictors for PM1. Chol: cholesterol. eGFR: estimated glomerular filtration rate. HDL: high-density lipoproteins. PM: predictive model. RBS: random blood sugar.



The PM3 model excludes the variable that showed the lowest importance, BMI. This model, when fitted using RCS with 5 knots, shows better performance using only the 5 predictors (age, RBS, cholesterol, eGFR, and non-HDL). The eGFR shows greater importance when fitted using RCS with 5 knots (>0.05) than when fitted with 3 knots (<0.05). The predictors' importance order for PM3 is shown in Figure 5. PM3 achieves an average accuracy of 74.73%, with a better confidence interval (95% CI 75%-78%). The calibration curve for PM3 is identical to that of PM1.

Figure 5. Order of importance of predictors for PM3. Chol: cholesterol. eGFR: estimated glomerular filtration rate. HDL: high-density lipoproteins. PM: predictive model. RBS: random blood sugar.



When using the PM2 model, the results show agreement with the results from PM1 for 93.27% (33,929/36,378) of predictions. The PM3 model with fewer predictors achieves a better performance and a similar percentage of predictions that are in agreement with the output from PM1 (33,937/36,378, 93.29%). Furthermore, the results show a strong degree of correlation among the probability outputs produced by the 3 models (r=0.97).

Discussion

Principal Results

Applying the method employed in the original study achieved an accuracy of 73% to 74% using a dataset collected from the Middle East, compared with 77% obtained from using a population from the United States in the original study. The findings from this replication study therefore confirm the

conclusion from the original study that this form of modeling can help with predicting the levels of HbA_{1c} in a blood test for nondiabetic patients using predictors extracted from EHR systems.

The order of importance obtained for the predictors used by the multiple logistic regression on our dataset is different from the order of importance produced in the original study. The order for the predictors using the KAIMRC dataset, from the most to the least importance, is RBS, age, eGFR, cholesterol, non-HDL, and BMI. Table 4 shows the importance rankings for the predictors obtained from the original study, as well as the rankings obtained from the 3 models used in this study.

Table 4. Predictors importance rankings.

Study	1st	2nd	3rd	4th	5th	6th	7th	8th
Original study	Age	BMI	RBS[a]	Race	Non-HDL[b]	Cholesterol	eGFR[c]	Smoking status
Replication study								
PM[d]1	RBS	Age	Cholesterol	Non-HDL	eGFR	BMI	N/A[e]	N/A
PM2	Age	RBS	Cholesterol	Non-HDL	BMI	eGFR	N/A	N/A
PM3	RBS	Age	eGFR	Cholesterol	Non-HDL	BMI (excluded)	N/A	N/A

[a]RBS: random blood sugar.

[b]HDL: high-density lipoproteins.

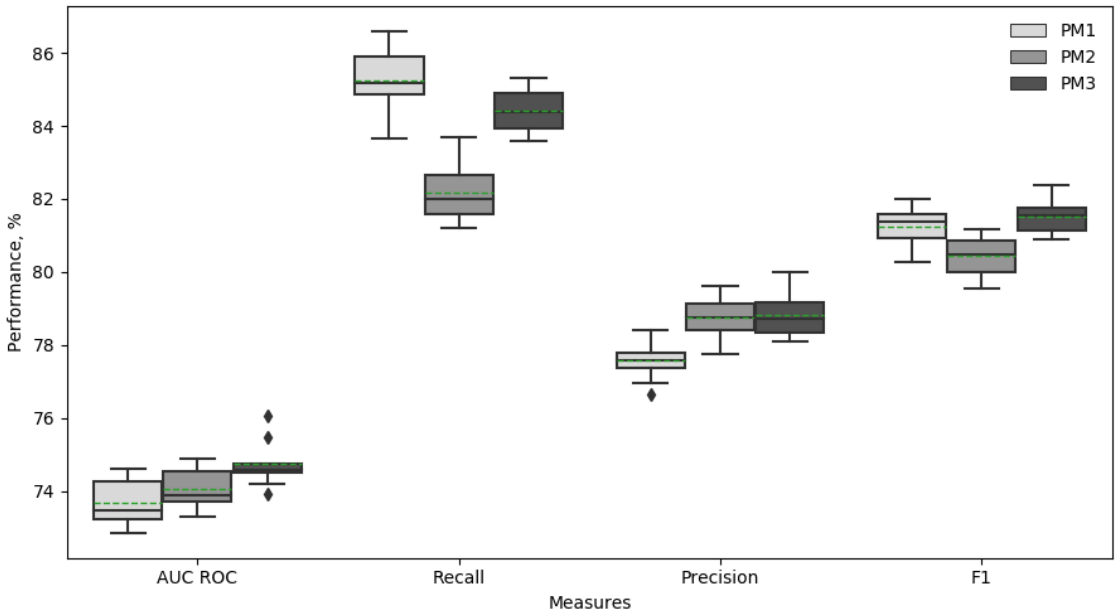
[c]eGFR: estimated glomerular filtration rate.

[d]PM: predictive model.

[e]N/A: not applicable.

BMI was one of the most important predictors in the population from the United States and demonstrated higher impact than the RBS and eGFR. However, it shows little importance for predicting the elevation level of HbA_{1c} in the KAIMRC population. Indeed, the simpler calculator with a reduced number of variables (after excluding BMI) is able to achieve better prediction abilities (refer to Multimedia Appendix 4 for details of the calculator). Figure 6 summarizes the 10-folds performance achieved using the reported measures for all models, and reveals that there is a consistent prediction trend for PM3, especially in the AUR ROC, which shows little variation between the folds.

Figure 6. Box plots of the reported measures for the models. AUC ROC: area under the receiver operating characteristic. PM: predictive model.



This replication study shows that the ranking of the variables is largely based on the dataset and the model used for prediction. Variables with low importance in the prediction of HbA_{1c} in one population may show greater or lesser importance when the model is applied on populations from different regions of the world. Interestingly, this can also happen when employing different predictive models and with different hyperparameters using the same population (for instance, eGFR shows higher importance when fitted to the model using RCS with 5 knots in PM3 than with 3 knots in PM1 and without RCS in PM2, as interpreted in Table 4).

Limitations and Future Work

We performed a differentiated replication using a population from a different region that was available to us. The 2 datasets have similar means and standard deviations for most of the

variables, such as age, cholesterol, and non-HDL, as described in Table 2. However, there is a significant difference in the body mass index and random blood sugar variables, and the dispersion is large for both variables.

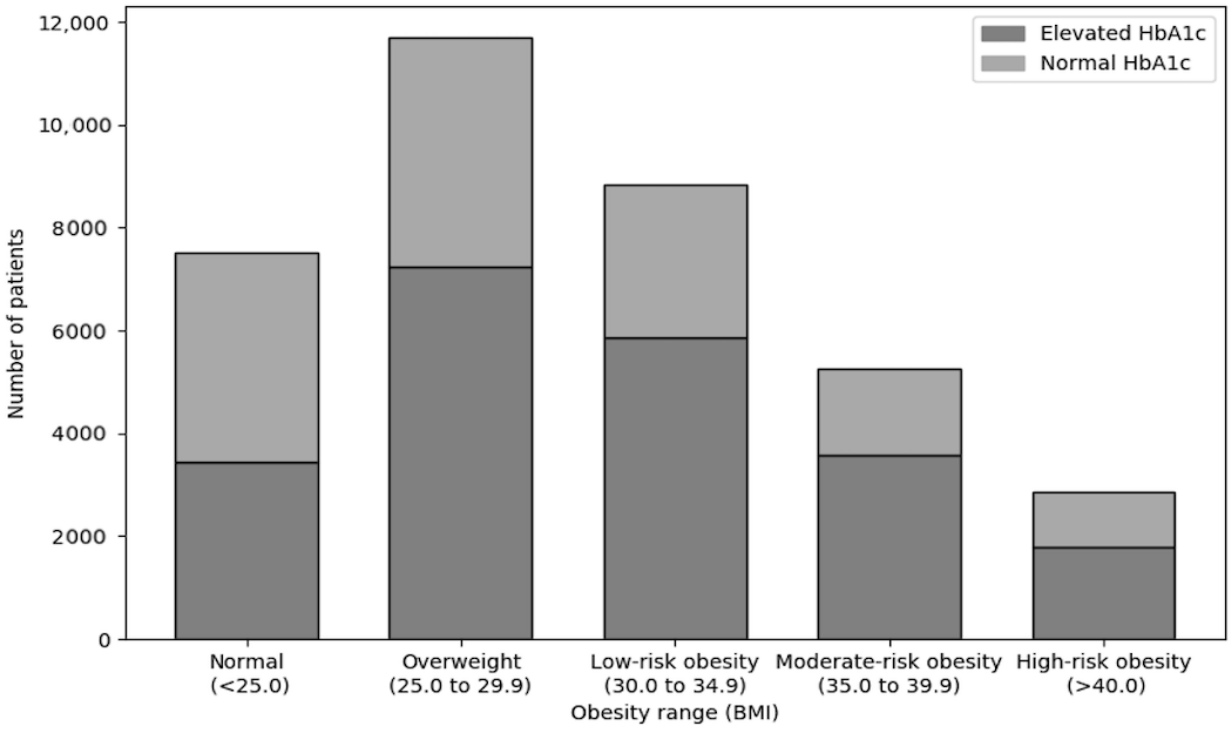
The sample size and class balance affect the learning behavior of the models [29]. The KAIMRC dataset is larger than the one used in the original study by 38%. The class balance is also different, with 26% of patients having elevated HbA_{1c} (≥5.7%) and 74% with normal HbA_{1c} (<5.7%) in the original study compared with 60.60% (22,046/36,378) with elevated HbA_{1c} (≥5.7%) and 39.40% (14,332/36,378) with normal HbA_{1c} (<5.7) in KAIMRC dataset.

Although the population represented in this study is less heterogeneous with regard to ethic groups, the size of the KAIMRC dataset is larger than the one used in the original study. The prevalence of diabetes is also larger, being a sample from the population of Saudi Arabia. In terms of prevalence of diabetes, Saudi Arabia was ranked by the World Health Organization as being the second highest in the Middle East and seventh highest in the world [30], with an 18.3% diabetes prevalence rate, according to the IDF, compared with 10.5% in the United States [31].

In the original study, the model performance was compared with the models developed by Baan et al [32] and Griffin et al [33], which used different datasets [34,35]. The main limitation in the comparison between the original study and the studies by Baan et al and Griffin et al is the absence of some variables that were used to create the calculators (refer to Multimedia Appendix 5 for details about the variables used in the corresponding studies). The same situation applies to this study, as the smoking status variable is missing in the KAIMRC dataset. The smoking prevalence in Saudi Arabia is between 2.4% to 52.3% among different age groups [36]. However, other missing predictors, such as genetic or lifestyle characteristics [37], which are difficult to collect and incorporate into the EHR systems, may help to explain the high rate of elevated levels of HbA_{1c} in the KAIMRC population.

After eliminating the variables that do not show significant impact on the prediction of HbA_{1c} in the KAIMRC population, the results indicate that different regions in the world can have different weightings of predictors for HbA_{1c} when using the approach of Wells et al. Although there are many studies that have demonstrated the relationship between diabetes prevalence and BMI [38], some studies have shown that the obesity prevalence in Asian countries does not relate to the diabetes prevalence. The risk of diabetes occurs in patients with a lower BMI in Asian countries compared with patients from European countries [39]. The prevalence of obesity in Asian countries is substantially less than in the United States, but Asian countries have a similar or higher prevalence of diabetes [40]. However, neither Yoon et al [39] nor Hu [40] identifies a relationship between nondiabetic patients with elevated levels of HbA_{1c} and obesity. Figure 7 visualizes the class distribution for the BMI variable for the KAIMRC dataset. The figure shows that elevation of HbA_{1c} exists with similar rates between low and high obesity ranges.

Figure 7. HbA_{1c} elevation for BMI ranges of King Abdullah International Medical Research Center patients. HbA_{1c}: glycated hemoglobin.



Advanced data mining techniques, such as deep machine learning models, are capable of finding hidden and complex correlations in large input spaces and datasets [41]. Recently, machine learning models have shown great success in many domains (eg, natural language processing, image segmentation, and object detection), but there is still a lack of studies that apply those models to the medical domain using EHR data [42]. As stated in the original study, maintaining security and privacy

for medical datasets is a challenging task. However, with advanced technologies in data privacy and protection, such as differential privacy and data anonymization techniques [43], it should be possible to minimize the security risk.

Conclusions

Replication studies provide an invaluable contribution to the validation, generalization, and continuation of scientific research. The differentiated replication presented in this study is aimed at validating the calculator used for predicting HbA_{1c} and evaluating the method used to create the mathematical equation by training the multiple logistic regression algorithm using EHR datasets. The evaluation was performed using a dataset collected from a different population. The original and replicated calculators employ associated predictors that are routinely collected and stored in hospital systems.

As explained in the "Introduction" section, this differentiated replication study used the same method to analyze a different population sample, with some differences in the form of the EHRs. As a replication, it was intended to investigate what changed and did not change in the outcomes.

What did not change appreciably was the accuracy of the results produced using this method, with an accuracy range of 73.6% to 74.7% in our study compared with 77% in the original study. The set of predictors (when these could be compared) also did not change. Thus, given that a close replication of the original study is unavailable, the differentiated replication does confirm that, despite the notable differences between the two datasets, the use of multiple logistic regression is able to provide good predictions of HbA_{1c} elevation levels.

What did change was the order of importance for the set of predictors used in the calculator. Thus, we can conclude that the use of multiple logistic regression for prediction does need to be tuned to the characteristics of the population being assessed. While we cannot wholly rule out the cause of this difference in importance being due to differences in the form of the EHRs, it seems more likely that the characteristics of the population were an important factor.

In terms of the role of replication itself, we would argue that this study demonstrates that while there is little difference in prediction accuracy when using multiple logistic regression with different populations (as might be expected), the influence of the different elements in the set of predictors is different. Due to that, we would argue that the generalization of simple statistical predictive models (calculators) is inappropriate. We suggest that creating advanced predictive models that can learn complex relationships using large multidimensional datasets may be a better way to exploit the increasing volumes of EHR data becoming available. Hence, further work will investigate applying advanced machine learning techniques to predict the elevation of HbA_{1c} using the KAIMRC dataset.

Acknowledgments

We would like to acknowledge the contribution of King Abdullah International Research Center (KAIMRC) for providing the dataset under the approved projects "Diabetes Early Warning System, Research Protocol SP14/042," "Finding the Common Related Diseases With Diabetes Using Data Mining Association Techniques, Research Protocol SP15/064," and extension project number RYD-17-417780-187503 to collect the newest dataset. We would also like to acknowledge the contribution by Professor Pali Hungin for providing feedback about the clinical aspects of the study.

Authors' Contributions

ZA was responsible for the designing, implementing, and building the statistical models. ZA and NAM were responsible for validating the models. ZA, DB, and NAM were responsible for the design of the replication study and for writing the manuscript. ZA and RA were responsible for extracting and describing the dataset. All authors participated in reviewing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Formulae for the calculated variables.
[PDF File (Adobe PDF File), 45 KB - medinform_v8i7e18963_app1.pdf]

Multimedia Appendix 2
Lab test and diagnostic codes.
[PDF File (Adobe PDF File), 72 KB - medinform_v8i7e18963_app2.pdf]

Multimedia Appendix 3
Units conversion formulae.
[PDF File (Adobe PDF File), 73 KB - medinform_v8i7e18963_app3.pdf]

Multimedia Appendix 4

PM3 Calculator details.
[PDF File (Adobe PDF File), 95 KB - medinform_v8i7e18963_app4.pdf]

Multimedia Appendix 5
Variables used in the studies.
[PDF File (Adobe PDF File), 58 KB - medinform_v8i7e18963_app5.pdf]

References

1. Cho N, Shaw J, Karuranga S, Huang Y, da Rocha Fernandes J, Ohlrogge A, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 2018 Apr;138:271-281. [doi: 10.1016/j.diabres.2018.02.023] [Medline: 29496507]
2. Wells BJ, Lenoir KM, Diaz-Garelli J, Futrell W, Lockerman E, Pantalone KM, et al. Predicting Current Glycated Hemoglobin Values in Adults: Development of an Algorithm From the Electronic Health Record. JMIR Med Inform 2018 Oct 22;6(4):e10780 [FREE Full text] [doi: 10.2196/10780] [Medline: 30348631]
3. Ogurtsova K, da Rocha Fernandes J, Huang Y, Linnenkamp U, Guariguata L, Cho N, et al. IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin Pract 2017 Jun;128:40-50. [doi: 10.1016/j.diabres.2017.03.024] [Medline: 28437734]
4. Beagley J, Guariguata L, Weil C, Motala AA. Global estimates of undiagnosed diabetes in adults. Diabetes Research and Clinical Practice 2014 Feb;103(2):150-160. [doi: 10.1016/j.diabres.2013.11.001] [Medline: 24300018]
5. Peterson KP, Pavlovich JG, Goldstein D, Little R, England J, Peterson CM. What is hemoglobin A1c? An analysis of glycated hemoglobins by electrospray ionization mass spectrometry. Clin Chem 1998 Sep;44(9):1951-1958. [Medline: 9732983]
6. Koenig RJ, Peterson CM, Jones RL, Saudek C, Lehrman M, Cerami A. Correlation of Glucose Regulation and Hemoglobin A in Diabetes Mellitus. N Engl J Med 1976 Aug 19;295(8):417-420. [doi: 10.1056/nejm197608192950804] [Medline: 934240]
7. International Expert Committee T. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. Diabetes Care 2009 Jul;32(7):1327-1334 [FREE Full text] [doi: 10.2337/dc09-9033] [Medline: 19502545]
8. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2010 Jan;33 Suppl 1:S62-S69 [FREE Full text] [doi: 10.2337/dc10-S062] [Medline: 20042775]
9. Ackermann RT, Cheng YJ, Williamson DF, Gregg EW. Identifying Adults at High Risk for Diabetes and Cardiovascular Disease Using Hemoglobin A1c. American Journal of Preventive Medicine 2011 Jan;40(1):11-17. [doi: 10.1016/j.amepre.2010.09.022] [Medline: 21146762]
10. Bonora E, Tuomilehto J. The pros and cons of diagnosing diabetes with A1C. Diabetes Care 2011 May;34 Suppl 2:S184-S190 [FREE Full text] [doi: 10.2337/dc11-s216] [Medline: 21525453]
11. Zhang X, Gregg EW, Williamson DF, Barker LE, Thomas W, McKeever Bullard K, et al. Response to Comment on: Zhang et al. A1C Level and Future Risk of Diabetes: A Systematic Review. Diabetes Care 2010;33:1665-1673. Diabetes Care 2011 Jan 26;34(2):e21-e21. [doi: 10.2337/dc10-2155] [Medline: 20587727]
12. Huang C, Iqbal U, Nguyen P, Chen Z, Clinciu DL, Hsu YE, et al. Using hemoglobin A1C as a predicting model for time interval from pre-diabetes progressing to diabetes. PLoS One 2014;9(8):e104263 [FREE Full text] [doi: 10.1371/journal.pone.0104263] [Medline: 25093755]
13. Ma W, Li H, Pei D, Hsia T, Lu K, Tsai L, et al. Variability in hemoglobin A1c predicts all-cause mortality in patients with type 2 diabetes. J Diabetes Complications 2012;26(4):296-300. [doi: 10.1016/j.jdiacomp.2012.03.028] [Medline: 22626873]
14. Khaw K, Wareham N, Bingham S, Luben R, Welch A, Day N. Association of hemoglobin A1c with cardiovascular disease and mortality in adults: the European prospective investigation into cancer in Norfolk. Ann Intern Med 2004 Sep 21;141(6):413-420. [doi: 10.7326/0003-4819-141-6-200409210-00006] [Medline: 15381514]
15. Pradhan AD, Rifai N, Buring JE, Ridker PM. Hemoglobin A1c predicts diabetes but not cardiovascular disease in nondiabetic women. Am J Med 2007 Aug;120(8):720-727 [FREE Full text] [doi: 10.1016/j.amjmed.2007.03.022] [Medline: 17679132]
16. McCarter RJ, Hempe JM, Chalew SA. Mean blood glucose and biological variation have greater influence on HbA1c levels than glucose instability: an analysis of data from the Diabetes Control and Complications Trial. Diabetes Care 2006 Feb;29(2):352-355. [doi: 10.2337/diacare.29.02.06.dc05-1594] [Medline: 16443886]
17. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ, A1c-Derived Average Glucose Study Group. Translating the A1C assay into estimated average glucose values. Diabetes Care 2008 Aug;31(8):1473-1478 [FREE Full text] [doi: 10.2337/dc08-0545] [Medline: 18540046]
18. Kazemi E, Hosseini S, Bahrampour A, Faghihimani E, Amini M. Predicting of trend of hemoglobin a1c in type 2 diabetes: a longitudinal linear mixed model. Int J Prev Med 2014 Oct;5(10):1274-1280 [FREE Full text] [Medline: 25400886]
19. Rose E, Ketchell D, Markova T. Clinical inquiries. Does daily monitoring of blood glucose predict hemoglobin A1c levels? J Fam Pract 2003 Jun;52(6):485-490. [Medline: 12791231]

20. Alhassan Z, Budgen D, Alessa A, Alshammari R, Daghstani T, Al moubayed N. Collaborative Denoising Autoencoder for High Glycated Haemoglobin Prediction. 2019 Presented at: International Conference on Artificial Neural Networks; Sep 17-19, 2019; Munich, Germany. [doi: 10.1007/978-3-030-30493-5_34]
21. Alhassan Z, Budgen D, Alshammari R, Daghstani T, McGough A, Al moubayed N. Stacked Denoising Autoencoders for Mortality Risk Prediction Using Imbalanced Clinical Data. 2018 Presented at: 17th IEEE International Conference on Machine Learning and Applications (ICMLA); Dec 17-20, 2018; Orlando, FL. [doi: 10.1109/icmla.2018.00087]
22. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-387. [doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4] [Medline: 8668867]
23. Gómez O, Juristo N, Vegas S, editors. Replications types in experimental disciplines. 2010 Presented at: ACM-IEEE international Symposium on Empirical Software Engineering and Measurement; Sep 16-17, 2010; Bolzano-Bozen, Italy. [doi: 10.1145/1852786.1852790]
24. Lindsay RM, Ehrenberg ASC. The Design of Replicated Studies. The American Statistician 1993 Aug;47(3):217. [doi: 10.2307/2684982]
25. Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer-Verlag; 2013.
26. Stone CJ. [Generalized Additive Models]: Comment. Statist Sci 1986 Aug;1(3):312-314. [doi: 10.1214/ss/1177013607]
27. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol 2012 Jun 20;12:82 [FREE Full text] [doi: 10.1186/1471-2288-12-82] [Medline: 22716998]
28. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
29. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl 2004 Jun;6(1):20-29. [doi: 10.1145/1007730.1007735]
30. Al Dawish MA, Robert AA, Braham R, Al Hayek AA, Al Saeed A, Ahmed RA, et al. Diabetes Mellitus in Saudi Arabia: A Review of the Recent Literature. Curr Diabetes Rev 2016;12(4):359-368. [doi: 10.2174/1573399811666150724095130] [Medline: 26206092]
31. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Centers for Disease Control and Prevention. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2020. URL: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf [accessed 2020-06-22]
32. Baan CA, Ruige JB, Stolk RP, Witteman JC, Dekker JM, Heine RJ, et al. Performance of a predictive model to identify undiagnosed diabetes in a health care setting. Diabetes Care 1999 Feb;22(2):213-219 [FREE Full text] [doi: 10.2337/diacare.22.2.213] [Medline: 10333936]
33. Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. Diabetes Metab Res Rev 2000;16(3):164-171. [doi: 10.1002/1520-7560(200005/06)16:3<164::aid-dmrr103>3.0.co;2-r] [Medline: 10867715]
34. Williams DRR, Wareham NJ, Brown DC, Byrne CD, Clark PMS, Cox BD, et al. Undiagnosed glucose intolerance in the community: the Isle of Ely Diabetes Project. Diabet Med 1995 Jan;12(1):30-35. [doi: 10.1111/j.1464-5491.1995.tb02058.x] [Medline: 7712700]
35. Kinmonth A, Spiegal N, Woodcock A. Developing a training programme in patient-centred consulting for evaluation in a randomised controlled trial; diabetes care from diagnosis in British primary care. Patient Educ Couns 1996 Oct;29(1):75-86. [doi: 10.1016/0738-3991(96)00936-6] [Medline: 9006224]
36. Bassiony MM. Smoking in Saudi Arabia. Saudi Med J 2009 Jul;30(7):876-881. [Medline: 19617999]
37. Elhadd TA, Al-Amoudi AA, Alzahrani AS. Epidemiology, clinical and complications profile of diabetes in Saudi Arabia: a review. Ann Saudi Med 2007;27(4):241-250 [FREE Full text] [doi: 10.5144/0256-4947.2007.241] [Medline: 17684435]
38. Boffetta P, McLerran D, Chen Y, Inoue M, Sinha R, He J, et al. Body mass index and diabetes in Asia: a cross-sectional pooled analysis of 900,000 individuals in the Asia cohort consortium. PLoS One 2011;6(6):e19930 [FREE Full text] [doi: 10.1371/journal.pone.0019930] [Medline: 21731609]
39. Yoon K, Lee J, Kim J, Cho JH, Choi Y, Ko S, et al. Epidemic obesity and type 2 diabetes in Asia. Lancet 2006 Nov 11;368(9548):1681-1688. [doi: 10.1016/S0140-6736(06)69703-1] [Medline: 17098087]
40. Hu FB. Globalization of diabetes: the role of diet, lifestyle, and genes. Diabetes Care 2011 Jun;34(6):1249-1257 [FREE Full text] [doi: 10.2337/dc11-0442] [Medline: 21617109]
41. Wischmeyer T, Rademacher T. Regulating Artificial Intelligence. Cham, Switzerland: Springer International Publishing; 2020.
42. Harerimana G, Kim JW, Yoo H, Jang B. Deep Learning for Electronic Health Records Analytics. IEEE Access 2019;7:101245-101259. [doi: 10.1109/access.2019.2928363]
43. Abadi M, Chu A, Goodfellow I, McMahan H, Mironov I, Talwar K. Deep learning with differential privacy. 2016 Presented at: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; Oct 24-28, 2016; Vienna, Austria. [doi: 10.1145/2976749.2978318]

Abbreviations

ADA: American Diabetes Association
AUR ROC: area under the receiver operating characteristic
eGFR: estimated glomerular filtration rate
EHR: electronic health record
HbA_{1c}: glycated hemoglobin
HDL: high-density lipoprotein
IDF: International Diabetes Federation
KAIMRC: King Abdullah International Medical Research Center
PM: predictive model
RBS: random blood sugar
RCS: restricted cubic splines
T2DM: type 2 diabetes mellitus

Edited by G Eysenbach; submitted 29.03.20; peer-reviewed by M Spiliopoulou; comments to author 26.04.20; revised version received 31.05.20; accepted 04.06.20; published 03.07.20.

Please cite as:
Alhassan Z, Budgen D, Alshammari R, Al Moubayed N
Predicting Current Glycated Hemoglobin Levels in Adults From Electronic Health Records: Validation of Multiple Logistic Regression Algorithm
JMIR Med Inform 2020;8(7):e18963
URL: https://medinform.jmir.org/2020/7/e18963
doi:10.2196/18963
PMID:32618575

Original Paper

Neural Network–Based Clinical Prediction System for Identifying the Clinical Effects of Saffron (Crocus sativus L) Supplement Therapy on Allergic Asthma: Model Evaluation Study

Seyed Ahmad Hosseini[1,2], DPhil; Amir Jamshidnezhad[1,3], DPhil; Marzie Zilaee[1], DPhil; Behzad Fouladi Dehaghi[4,5], DPhil; Abbas Mohammadi[4,5], DPhil; Seyed Mohsen Hosseini[3], MSc

[1]Nutrition and Metabolic Diseases Research Center, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

[2]Department of Nutrition, Faculty of Allied Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

[3]Department of Health Information Technology, Faculty of Allied Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

[4]Environmental Technologies Research Center, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

[5]Department of Occupational Health, School of Public Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

Corresponding Author:
Amir Jamshidnezhad, DPhil
Department of Health Information Technology
Faculty of Allied Medical Sciences
Ahvaz Jundishapur University of Medical Sciences
Golestan Boulevard, Esfand St. Faculty of Allied Medical Sciences
Ahvaz
Iran
Phone: 98 9166126106
Email: dr.jamshidnejad@gmail.com

Abstract

Background: Asthma is commonly associated with chronic airway inflammation and is the underlying cause of over a million deaths each year. *Crocus sativus L*, commonly known as saffron, when used in the form of traditional medicines, has demonstrated anti-inflammatory effects which may be beneficial to individuals with asthma.

Objective: The objective of this study was to develop a clinical prediction system using an artificial neural network to detect the effects of *C sativus L* supplements on patients with allergic asthma.

Methods: A genetic algorithm–modified neural network predictor system was developed to detect the level of effectiveness of *C sativus L* using features extracted from the clinical, immunologic, hematologic, and demographic information of patients with asthma. The study included data from men (n=40) and women (n=40) individuals with mild or moderate allergic asthma from 18 to 65 years of age. The aim of the model was to estimate and predict the level of effect of *C sativus L* supplements on each asthma risk factor and to predict the level of alleviation in patients with asthma. A genetic algorithm was used to extract input features for the clinical prediction system to improve its predictive performance. Moreover, an optimization model was developed for the artificial neural network component that classifies the patients with asthma using *C sativus L* supplement therapy.

Results: The best overall performance of the clinical prediction system was an accuracy greater than 99% for training and testing data. The genetic algorithm–modified neural network predicted the level of effect with high accuracy for anti–heat shock protein (anti-HSP), high sensitivity C-reactive protein (hs-CRP), forced expiratory volume in the first second of expiration (FEV_1), forced vital capacity (FVC), the ratio of FEV_1/FVC, and forced expiratory flow ($FEF_{25\%-75\%}$) for testing data (anti-HSP: 96.5%; hs-CRP: 98.9%; FEV_1: 98.1%; FVC: 97.5%; FEV_1/FVC ratio: 97%; and $FEF_{25\%-75\%}$: 96.7%, respectively).

Conclusions: The clinical prediction system developed in this study was effective in predicting the effect of *C sativus L* supplements on patients with allergic asthma. This clinical prediction system may help clinicians to identify early on which clinical factors in asthma will improve over the course of treatment and, in doing so, help clinicians to develop effective treatment plans for patients with asthma.

(JMIR Med Inform 2020;8(7):e17580) doi:10.2196/17580

KEYWORDS

asthma; machine learning; clinical predictor system; neural networks; supplement therapy; saffron; Crocus sativus L

Introduction

Asthma is a heterogeneous disease and usually coincident with chronic airway inflammation. Asthma can be diagnosed based upon the patient's history of respiratory symptoms such as wheezing, chest tightness, shortness of breath, and cough which may vary among the patient population in terms of intensity, time, and decreased expiratory airflow.

Allergic asthma is the easiest phenotype of asthma to diagnose with symptoms usually becoming apparent in childhood [1]. Globally, the prevalence of asthma in adults is estimated to be 4.3% [1]; in 2012, an estimated 300 million adults suffered from asthma, and by 2025, this figure will increase to 400 million. Moreover, the annual mortality rate of asthma is estimated to be about 250,000 [2]. In Iran, the prevalence of asthma is around 5.5% [3]. Every year, according to the World Health Organization (WHO), 15 million disability-adjusted life-years are lost because of the disease [4]. Those who develop asthma often have allergic conditions or a family history of allergic conditions such as eczema, food allergies, drug allergies, or allergic rhinitis. These patients normally respond well to inhaled corticosteroid treatments [2]. Like mast cells, different immunologic cells such as eosinophils, lymphocytes, and neutrophils have a role in the process of airway inflammation [5]. Genetic factors are among the risk factors of asthma, in addition to environmental factors such as exposure to allergens which may also exacerbate asthma symptoms and access to health care services [2].

There is a need for health care providers and governments to work collectively to improve control of asthma symptoms [2]. In traditional medicine, *Crocus sativus L*, which is more commonly known as saffron, has been used as a treatment for heart disease, depression, stress, and sleep disorders [5]. *C sativus L* possesses antioxidant [6] and anti-inflammatory properties [7]. Its active components—safranal and crocin—have demonstrated beneficial anti-inflammatory and antioxidant effects [5].

Several studies [6,8,9] have reported the effects of *C sativus L* on asthmatic patients. Zilaee et al [6] studied, in a randomized clinical trial, the effects of *C sativus L* supplements on clinical symptoms, Asthma Severity Score, blood pressure, and lipid profiles of patients with mild or moderate persistent allergic asthma. Although these studies showed aggregate effects, it was not possible to predict rare and serious effects for individuals.

Recommender systems, also known as recommendation engines, are used in online personalized predictive models and have been increasingly implemented in many areas of application to extract useful information from data; however, most of the available approaches that rely upon traditional statistical outcomes are unable to extract crucial knowledge [10,11] such as severity reduction estimates for patients with asthma [6]. Therefore, recommender systems can overlook significant effects of *C sativus L* supplements in patients with allergic asthma.

Recently, researchers have begun to estimate the effectiveness of clinical medicine using machine learning methods [12]. Machine learning techniques can be used to approximate the treatment effects of medicines [13,14]. An important application of machine learning in medicine is the development of automated risk-prediction algorithms to guide clinical care [15]. These algorithms can be used to integrate and interpret complex biomedical and health care data in scenarios where traditional statistical methods may not work [7].

To address the current limitations, a clinical prediction system based upon machine learning algorithms was developed to estimate the level of effect of *C sativus L* supplements in patients with allergic asthma. To classify clinical improvement in patients, we developed a model that determines the potential effect of *C sativus L* supplements on individual patients with asthma by extracting the factors with the greatest effect from clinical features, hematologic features, anti-inflammatory features, and Asthma Severity Score.

Methods

Data Description

To develop and evaluate a genetic algorithm–modified neural network model, we used a dataset [6] containing data on men (n=40) and women (n=40) with asthma who ranged in age from 18 to 65 years and who received *C sativus L* supplements. Using diagnostic criteria from the Global Initiative for Asthma, these patients had been diagnosed with mild or moderate allergic asthma by a pulmonologist in May 2017 or October 2017 and were recruited from the outpatient clinic at Imam Khomeini Hospital in Ahvaz, Iran.

Participants were asked to take one oral capsule containing 50 mg of dried *C sativus L* stigma (from the Faculty of Pharmacy at Ahvaz Jundishapur University of Medical Sciences) twice daily at 12 hour intervals. The *C sativus L* stigma was procured from Estahban, Fars Province, Iran (Herbarium code: JPS018118). The capsules contained 50 mg of dried saffron stigma and starch (as fillers). Each participant was asked to fill out questionnaires about asthma clinical symptoms and to have a sample of blood drawn. Additionally, the participants were interviewed about their socio-demographic status, job, smoking, medical background, and medication.

Clinical Symptoms and Critical Factors

Clinical symptoms (frequency of shortness of breath during the day, frequency of shortness of breath during the night, limitations on activity, frequency of salbutamol inhaler use, and sleep problems caused by asthma symptoms) were recorded at the preintervention and postintervention. Tests for hematologic (eosinophil and basophil counts) and anti-inflammatory factors (anti-HSP: anti–heat shock protein; and hs-CRP: high sensitivity C-reactive protein), and spirometry tests (FEV_1: forced expiratory volume in the first second of expiration; FVC: forced vital capacity; FEV_1/FVC ratio; and $FEF_{25\%-75\%}$: midphase forced expiratory flow) were conducted preintervention and

postintervention. Demographic information (age, BMI, gender, weight, and smoking history) were recorded at preintervention. Table 1 lists the input parameters, their units of measurement, and their factor type.

Table 1. Model inputs.

Input parameters	Units of measure	Factor type
FEV_1[a]	Liter	clinical
FVC[b]	Liter	clinical
FEV_1/FVC	Liter	clinical
$FEF_{25\%-75\%}$[c]	L/minute	clinical
Shortness of breath during the day	frequency per day	clinical
Shortness of breath during the night	frequency per day	clinical
Waking up due to asthma symptoms	frequency per day	clinical
Activity limitation	frequency per day	clinical
Salbutamol inhaler use	frequency per day	clinical
anti-HSP70[d]	ng/mL	anti-inflammatory
hs-CRP[e]	ng/mL	anti-inflammatory
Eosinophil	number/ μL	hematologic
Basophil	number/ μL	hematologic
Age	years	demographic
Smoking history	yes or no	demographic
BMI[f]	kg/m^2	demographic
Gender	male or female	demographic
Weight	kilogram	demographic

[a]FEV_1: forced expiratory volume in 1 s.

[b]FVC: forced vital capacity.

[c]$FEF_{25\%-75\%}$: forced expiratory flow.

[d]anti-HSP: anti-heat shock protein.

[e]hs-CRP: high sensitivity C-reactive protein.

[f]BMI: body mass index.

Genetic Algorithm–Modified Neural Network Model

Overview

Asthma improvement was treated as an information retrieval problem; therefore, our model was developed to determine the relationship between *C sativus L* and four types of factors: clinical, anti-inflammatory, hematological, and demographic measures.

An artificial neural network (ANN) machine learning model [16] was developed using MATLAB (version R2018b; MathWorks Inc) software and was used to classify patients with asthma. As a preliminary processing step, selection criteria were used to identify which parameters were discriminant predictors in order to enhance algorithm performance by eliminating those that were redundant or irrelevant [15]. To select these parameters, a genetic algorithm was used to assess subsets of parameters according to each parameter's contribution to diagnostic performance; thus, discriminant predictors were selected as factors by their effect on the diagnostic performance of the model. The genetic algorithm was applied to find the optimal structure of the ANN model. Patients with allergic asthma who received *C sativus L* supplement were classified into 5 classes.

Artificial Neural Network

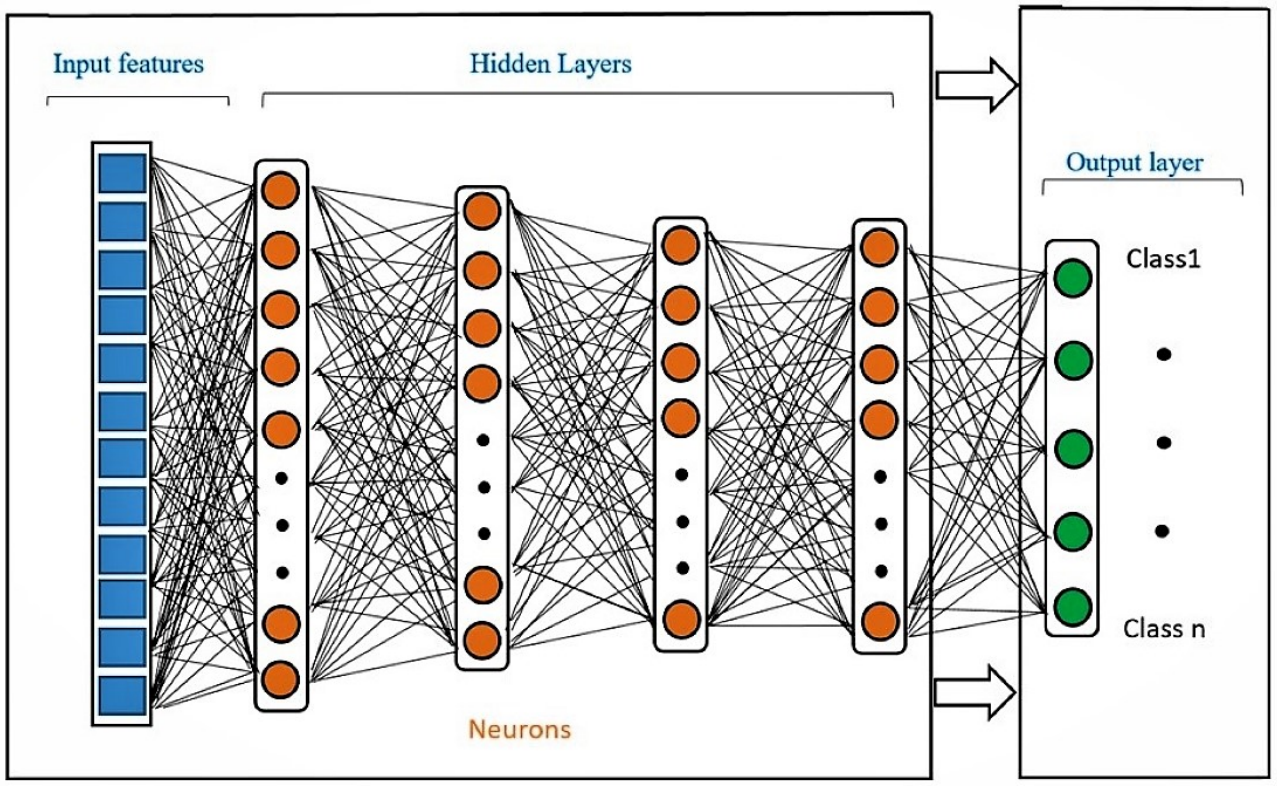
It is possible to solve many problems using an ANN since they are capable of computing any function which is computable. These networks are mostly suitable for solving problems that can tolerate specific levels of error. ANNs are built from several layers of interconnected nodes which are called neurons. In a typical feedforward neural network, there is at least one input layer, one hidden layer, and one output layer. The number of nodes in the input layer corresponds to the number of input features; the features are analogous to the covariates or independent variables that are incorporated into a linear regression model. The output nodes represent predictions or classifications. Backpropagation algorithms in an ANN are able to train the model using teacher-based supervised learning [17]. Testing performance from backpropagation is not always

satisfactory, even if training performance demonstrates high accuracy [18].

Although, many different forms of ANN exist, this study uses ANNs built from several hidden layers which are often used in nonparametric problems. Connection weights between each neighboring layer are continuously updated so that output values approach the targeted values.

In designing feedforward neural network topology for prediction, the components which must be considered are input layer configuration, hidden layer configuration, and output layer configuration as well as the model's training methodology. In practice, this architecture is determined by experimentation [15]. There is a direct relationship between the input and hidden layers which operate in conjunction to apply weights to inputs and that result in new outputs (Figure 1). Eventually, the output layer classifies or predicts the outcome of the process based upon the transmitted values. The advantage of an ANN is that the network is comprised of multiple nonlinear levels which makes the ANN capable of representing highly varying functions [18]. ANNs can be used to determine complex patterns in data and may be applied in medical fields.

Figure 1. General structure of feedforward multilayer neural network.



Genetic Algorithm

To find the optimal architecture of the multilayer ANN model, the input parameters, neurons, hidden layers, and learning functions were optimized using a genetic algorithm model. The genetic algorithm is a well-known optimization technique that may be categorized as an evolutionary method using biological process [19]. Genetic algorithms have been demonstrated to be reliable and robust in many medical applications [20-22]. A genetic algorithm that included chromosome reproduction, crossover, and mutation heuristic processes was implemented.

The genetic algorithm–modified neural network predictive model is presented in Figure 2. The best architecture for the neural network was found by iteration of the model using the genetic algorithm that is shown in Figure 3. The genetic algorithm was used to find the most effective features to estimate the level of variation of every parameter in the patients with allergic asthma after *C sativus L* supplement therapy. Iteration was terminated when the optimal architecture was reached in terms of effective input parameters (factors) as well as hidden layers, neurons, and learning function to find the highest accuracy for prediction of effects. The accuracy rate achieved from the optimal architecture in each generation was the best fitness value. Optimization of the model occurred in two phases: (1) estimation of level of variation of each factor, separately, and (2) classification of the level of alleviation of the severity of asthma. In the first prediction process, the genetic algorithm–modified neural network system estimated the rate of variation of factors and in the second phase, the patients were classified into 5 groups from low to high level of effect of *C sativus L* supplement. The optimized neural network model predicted the level of change for each factor postintervention and classified patients into 5 groups by asthma condition.

Figure 2. The architecture of the genetic algorithm–modified neural network system for predicting *C sativus L* supplement effects on patients with allergic asthma. Anti-HSP: anti-heat shock protein; hs-CRP: high sensitivity C-reactive protein; FEV_1: forced expiratory volume in 1 s; FVC: forced vital capacity; $FEF_{25\%-75\%}$: forced expiratory flow; BMI: body mass index.

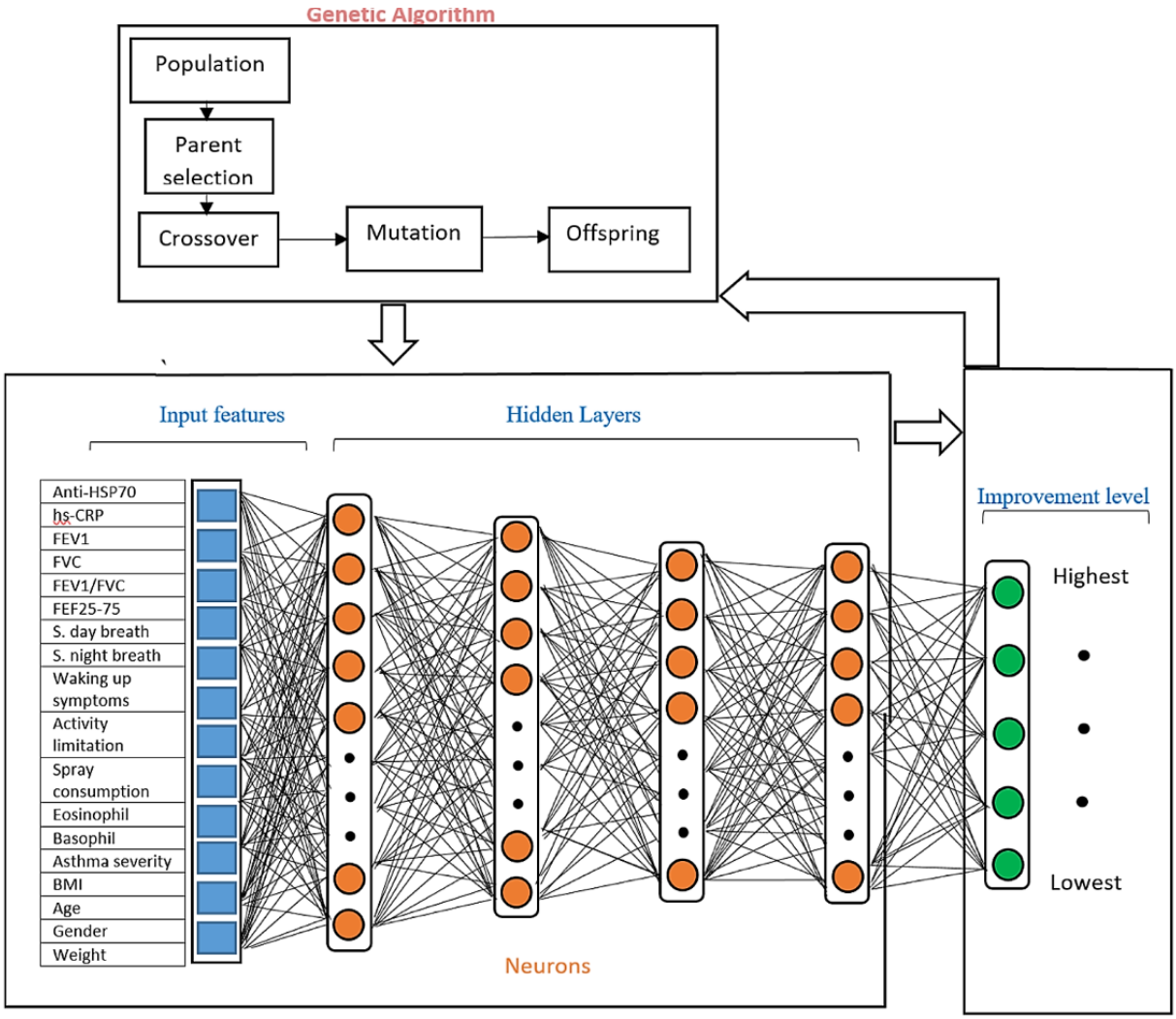
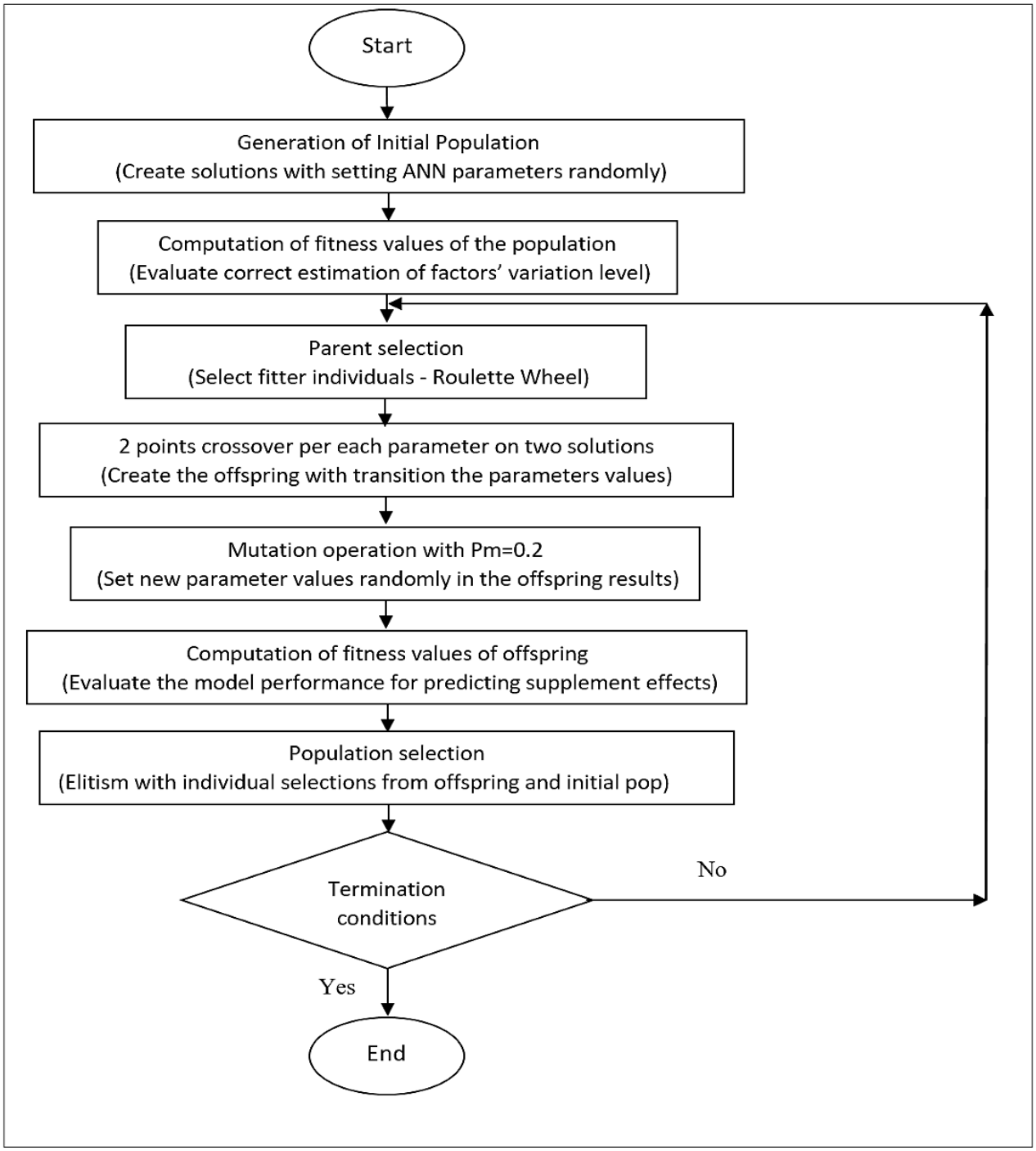


Figure 3. The process of the proposed genetic algorithm. ANN: artificial neural network.



Performance and Validation

Model Performance Metrics

Mean squared error (MSE) is a commonly used error function in ANN training to evaluate the efficiency of the model. This function calculates the mean error then sends the error back to the nodes [23]. The equation for MSE and accuracy are defined as $MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$ where N is the total number of outputs, f_i is the desired output value for each output, i (i=1, 2,…, N), and y_i is the neural network output value for each output, i, and as $accuracy = 1 - MSE$.

Cross-Validation

The dataset was randomly subdivided into training (90% of available data) and testing datasets (10% of available data) using the Monte Carlo method for cross-validation (also known as repeated random subsampling) [24]. This method generated multiple random divisions of data into training and testing in 10 experiments. Furthermore, each experiment was iterated 10 times to validate of the model.

The protocol of this study was approved by the Medical Ethics Committee at Ahvaz Jundishapur University of Medical Sciences (IR.AJUMS.REC.1395.810, IR.AJUMS.REC.1398.880).

Results

The best fitness value that was achieved for each output factor is shown in Table 2. Model performance is shown in Figures 4-7. The results were achieved with 100 epochs for training process.

Table 2. Best training performance at epoch 100.

Factor[a]	Fitness value
FEV_1[a]	0.007
FVC[b]	0.003
FEV_1/FVC	0.006
$FEF_{25\%-75\%}$[c]	0.002
Shortness of breath during the day	0.002
Shortness of breath during the night	0.001
Waking up due to asthma symptoms	0.004
Activity limitation	0.001
Salbutamol inhaler use	0.002
anti-HSP70[d]	0.003
hs-CRP[e]	0.006
Eosinophils	0.001
Basophils	0.002

[a]FEV_1: forced expiratory volume in 1 s.

[b]FVC: forced vital capacity.

[c]$FEF_{25\%-75\%}$: forced expiratory flow.

[d]anti-HSP: anti-heat shock protein.

[e]hs-CRP: high sensitivity C-reactive protein.

Figure 4. Training performance of the genetic algorithm–modified neural network for *C sativus L* supplement effects on the anti-inflammatory factors: anti-HSP70 (top left, anti-heat shock protein); hs-CRP (high sensitivity C-reactive protein, top right); and hematologic factors: basophil (bottom left); eosinophil (bottom right).

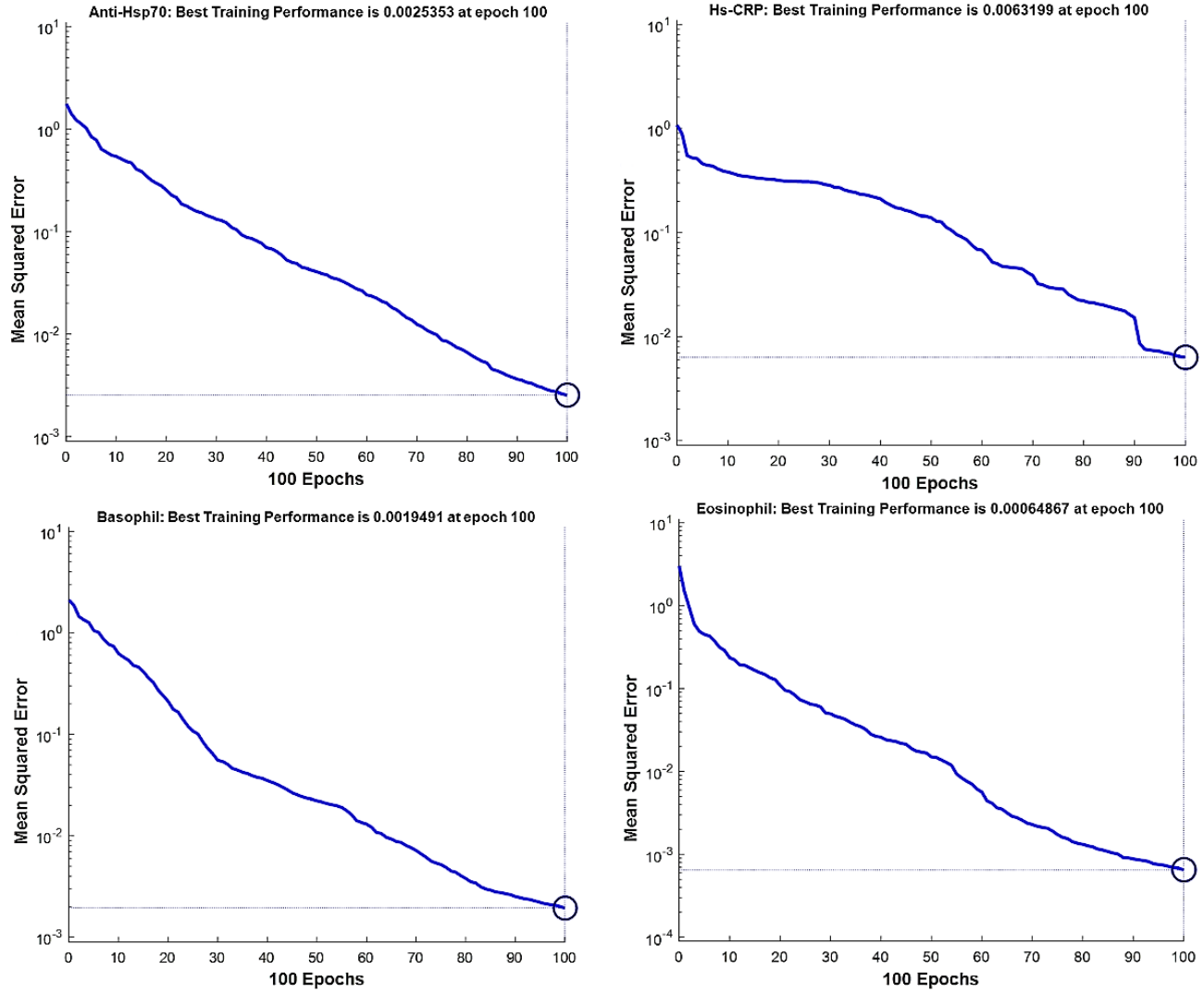


Figure 5. Training performance of the genetic algorithm–modified neural network for *C sativus L* supplement effects on the clinical factors: FEV_1/FVC ratio (top left); FVC (top right); FEF-25 (forced expiratory flow, bottom left); FEV (forced expiratory volume, bottom right).

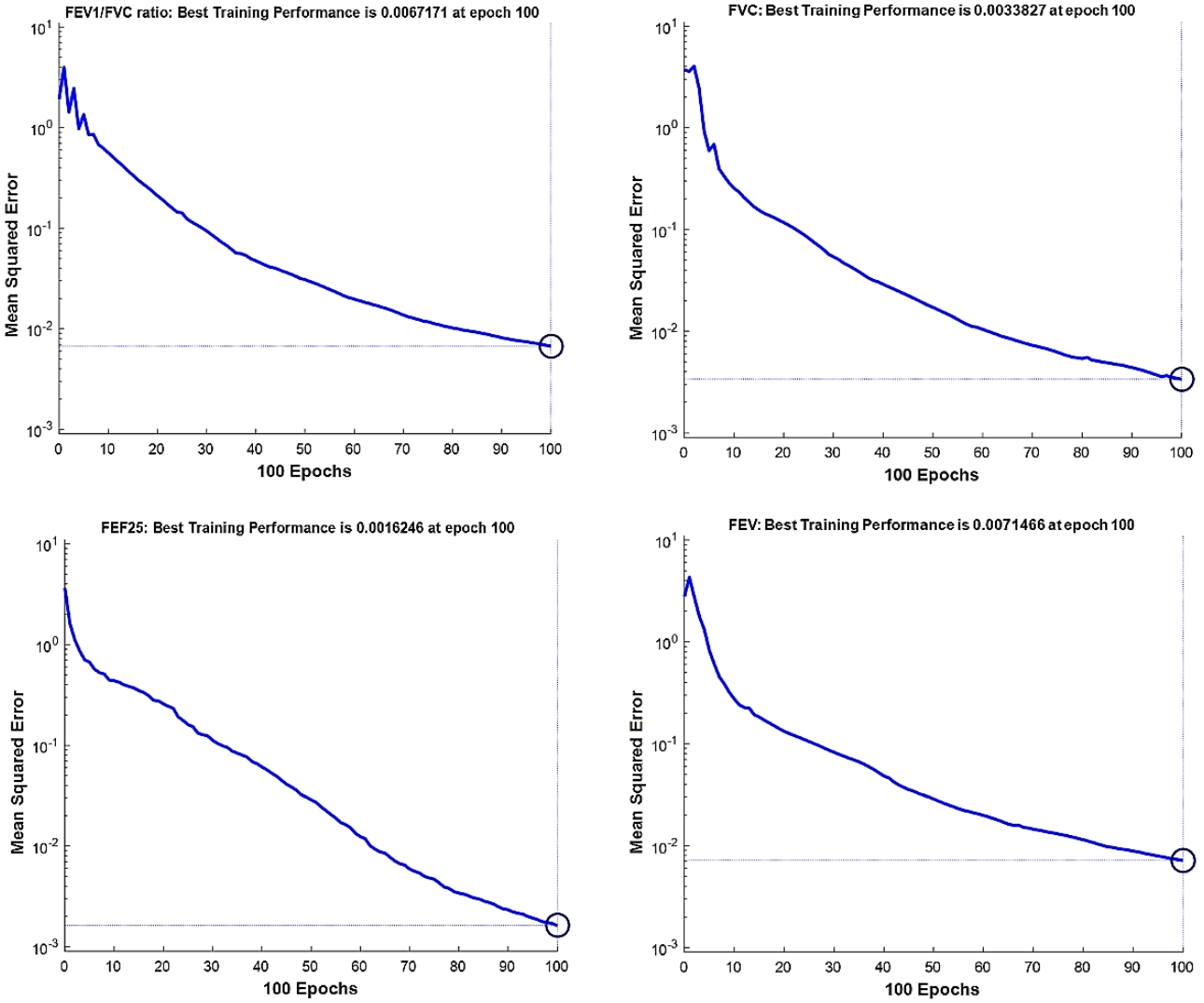


Figure 6. Training performance of the genetic algorithm–modified neural network for *C sativus L* supplement effects on the asthma clinical symptoms: shortness of breath in the night (top left); shortness of breath during the day (top right); waking up due to asthma symptoms (bottom).

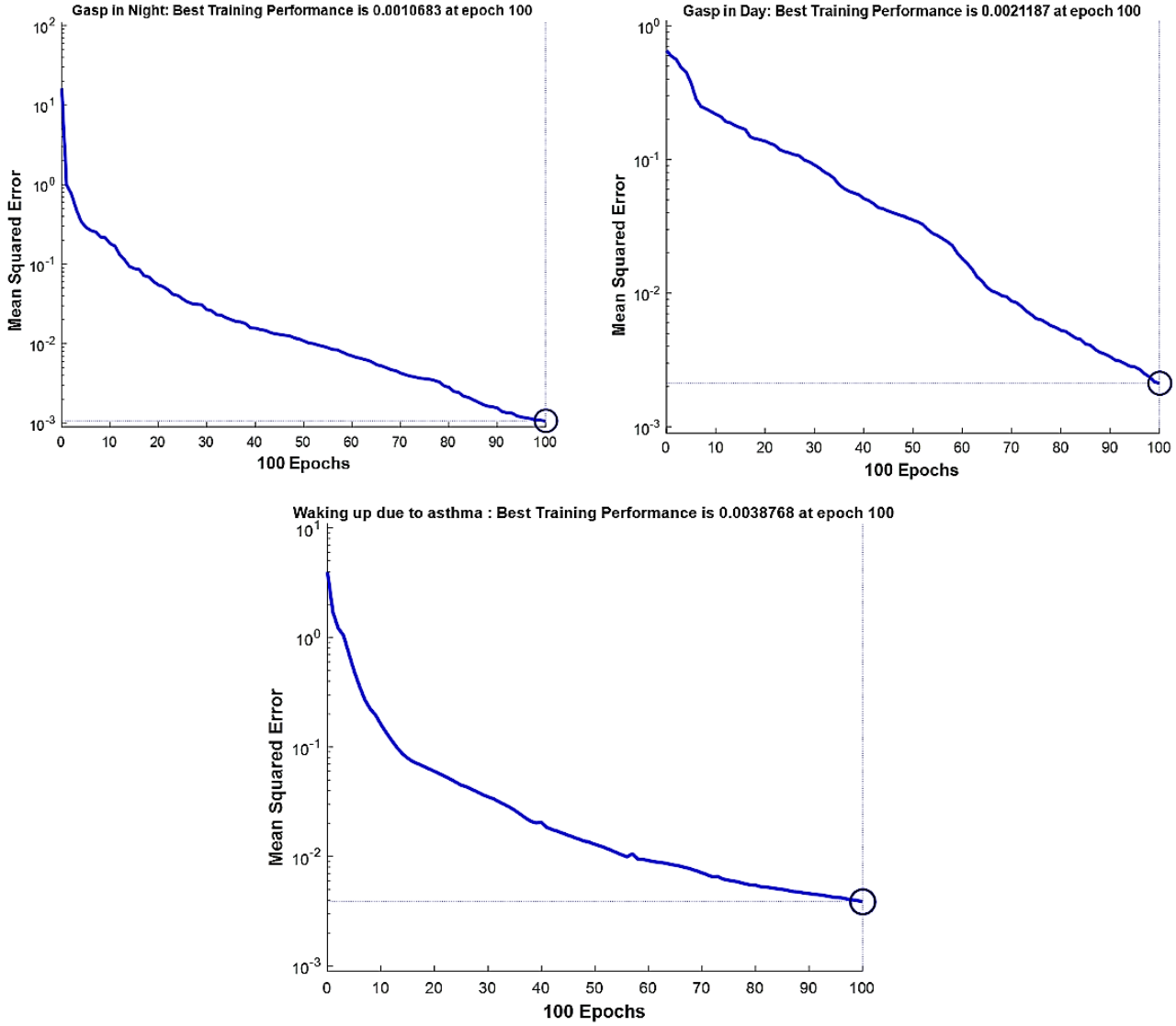
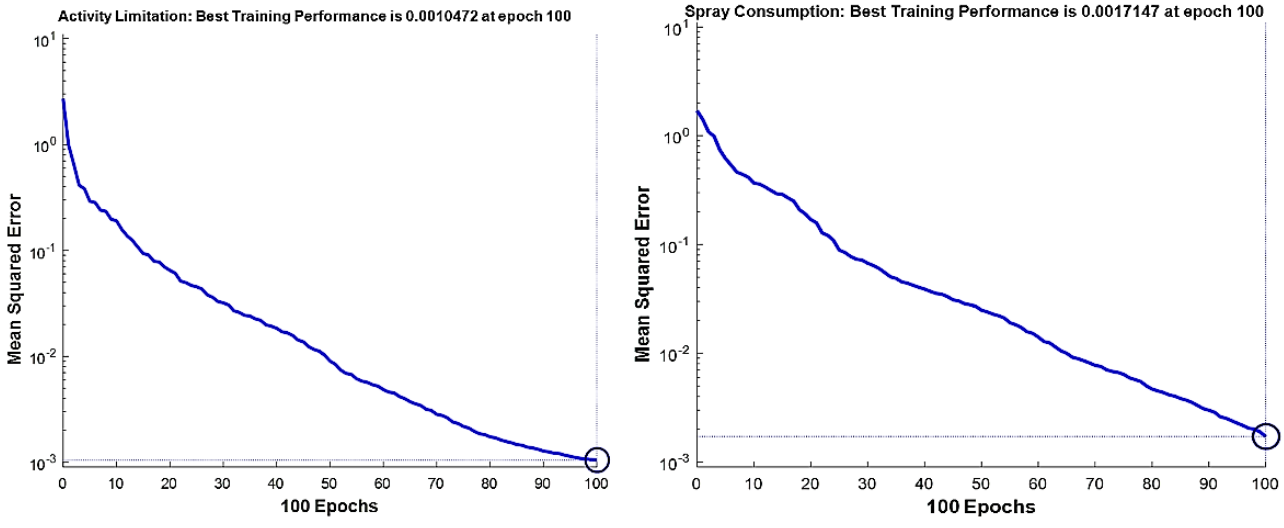


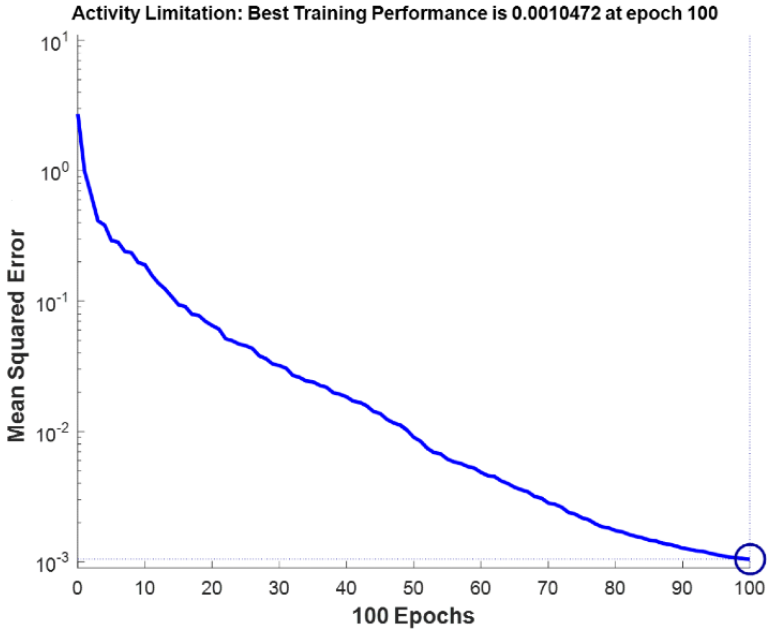
Figure 7. Training performance of the genetic algorithm–modified neural network for *C sativus L* supplement effects on the asthma clinical symptoms: activity limitation (left), salbutamol inhaler use (right).



As shown in Figure 4-7, the model's training performance reached the acceptable fitness value with the MSE lower than 5% for each factor. Furthermore, Figure 8 shows the training performance of the overall model which reached a fitness level

of 3.3322×10^{-9}. This result illustrates an accuracy greater than 99% for training. Table 3 shows the most effective factors that were found by the genetic algorithm. The selected parameters were used as input in the genetic algorithm–modified neural network to predict the effect on each clinical, hematologic, or anti-inflammatory parameter. FEV_1/FVC ratio, waking up due to asthma symptoms, hs-CRP, eosinophil, basophil, weight, and smoking history were selected most frequently by the genetic algorithm process to estimate the level of alleviation after *C sativus L* supplement therapy. Table 4 shows the optimized architecture of the model for the test set (hidden layers, neurons, training and fitting functions), and estimated rate of risk after supplement therapy for each factor are presented. According to Table 4, the predictions for waking up due to asthma symptoms and basophil with the MSE of 0.004 and 0.045 showed the highest and lowest accuracy, respectively, among the factors in the testing phase. A pattern recognition network was selected by the genetic algorithm for most experiments (patternnet in MATLAB). Furthermore, conjugate gradient backpropagation with Polak-Ribière updates (traincgp in MATLAB) was frequently selected for training in the genetic algorithm optimization process.

Figure 8. Training performance of the genetic algorithm–modified neural network for total classification of *C sativus L* supplement effects on patients with allergic asthma.



According to Table 4, the best accuracy was for waking due to asthma symptoms; the model classified the patients in terms of level of alleviation with an accuracy greater than 99% (MSE=0.0004) in testing. For this prediction model, 6 factors (FEV_1/FVC ratio, frequency of waking due to asthma symptoms, hs-CRP, eosinophil, basophil, weight, smoking condition) out of 18 input parameters were identified and used to classify the patients in five groups from low to high level of effect.

Table 3. Genetic algorithm feature selection results for prediction of risk factors.

Inputs selected by the genetic algorithm[a]	Factor
FEV_1[a]/FVC[b], $FEF_{25\%-75\%}$[c], activity limitation, salbutamol inhaler use, anti-HSP70[d], hs-CRP[e], eosinophil, gender, weight, smoking condition	FEV_1
FVC, FEV_1/FVC, activity limitation, hs-CRP, basophil, age, gender, weight	FVC
FVC, $FEF_{25\%-75\%}$, shortness of breath during the day, shortness of breath during the night, activity limitation, salbutamol inhaler use, hs-CRP, basophil, age, weight, smoking condition	FEV_1/FVC
FEV_1, FVC, $FEF_{25\%-75\%}$, salbutamol inhaler use, anti-HSP70, basophil, gender, BMI[f]	$FEF_{25\%-75\%}$
FVC, $FEF_{25\%-75\%}$, shortness of breath during the day, shortness of breath during the night, activity limitation, salbutamol inhaler use, basophil	shortness of breath during the day
FEV_1, FEV_1/FVC, $FEF_{25\%-75\%}$, shortness of breath during the night, waking up due to asthma symptoms, salbutamol inhaler use, anti-HSP70, age, smoking condition	shortness of breath during the night
Shortness of breath during the day, shortness of breath during the night, waking up due to asthma symptoms, anti-HSP70, eosinophil, weight	waking up due to asthma symptoms
FEV_1, FVC, FEV_1/FVC, shortness of breath during the night, waking up due to asthma symptoms, activity limitation, salbutamol inhaler use, eosinophil, basophil, age, smoking condition	activity limitation
FEV_1, shortness of breath during the night, activity limitation, salbutamol inhaler use, weight, BMI	salbutamol inhaler use
FVC, waking up due to asthma symptoms, salbutamol inhaler use, hs-CRP, eosinophil, basophil, age, gender, weight	anti-HSP70
FEV_1, FVC, $FEF_{25\%-75\%}$, shortness of breath during the day, waking up due to asthma symptoms, activity limitation, anti-HSP70, eosinophil, basophil, weight, BMI, smoking condition	hs-CRP
FEV_1, waking up due to asthma symptoms, anti-HSP70, hs-CRP, eosinophil, basophil, BMI, age, gender, weight	eosinophil
FEV_1, FVC, shortness of breath during the night, waking up due to asthma symptoms, salbutamol inhaler use, anti-HSP70, hs-CRP, eosinophil, basophil, gender, weight, BMI	basophil
FEV_1/FVC, waking up due to asthma symptoms, hs-CRP, eosinophil, basophil, weight, smoking condition	total factors

[a]FEV_1: forced expiratory volume in 1 s.

[b]FVC: forced vital capacity.

[c]$FEF_{25\%-75\%}$: forced expiratory flow.

[d]anti-HSP: anti-heat shock protein.

[e]hs-CRP: high sensitivity C-reactive protein.

[f]BMI: body mass index.

Table 4. Genetic algorithm–modified neural network optimization results for the best testing prediction of *C sativus L* effects.

Estimated risk factor	Optimized hidden layers	Training function[a]	Fitting model[b]	MSE[c]
FEV_1[d]	2 layers: 16 × 25 neurons	trainrp	patternnet	0.019
FVC[e]	2 layers: 26 × 25 neurons	trainrp	patternnet	0.025
FEV_1/FVC	2 layers: 16 × 15 neurons	trainrp	patternnet	0.029
$FEF_{25\%-75\%}$[f]	2 layers: 25 × 19 neurons	traincgp	patternnet	0.033
Shortness of breath during the day	2 layers: 12 × 7 neurons	traincgb	patternnet	0.010
Shortness of breath during the night	2 layers: 18 × 8 neurons	traincgb	feedforwardnet	0.005
Waking up due to asthma symptoms	2 layers: 22 × 6 neurons	trainrp	patternnet	0.004
Activity limitation	2 layers: 9 × 24 neurons	traincgf	patternnet	0.009
Salbutamol inhaler use	2 layers: 16 × 28 neurons	traincgb	patternnet	0.030
anti-HSP70[g]	2 layers: 27 × 17 neurons	traincgp	patternnet	0.035
hs-CRP[h]	2 layers: 10 × 3 neurons	trainoss	patternnet	0.012
Eosinophil	2 layers: 13 × 26 neurons	traincgb	patternnet	0.025
Basophil	2 layers: 26 × 8 neurons	traincgp	patternnet	0.045
Total	2 layers: 12 × 19 neurons	trainlm	patternnet	0.0004

[a]MATLAB training functions where trainrp is resilient backpropagation; traincgp, traincgb, and traincgf are conjugate gradient backpropagation with Polak-Ribière updates, Powell-Beale restarts, and Fletcher-Reeves updates, respectively; trainoss uses the one-step secant method; and trainlm uses Levenberg-Marquardt optimization.

[b]MATLAB fitting models where patternnet is a pattern recognition network, and feedforwardnet is a feedforward network.

[c]MSE: mean squared error.

[d]FEV_1: forced expiratory volume in 1 s.

[e]FVC: forced vital capacity.

[f]$FEF_{25\%-75\%}$: forced expiratory flow.

[g]anti-HSP: anti-heat shock protein.

[h]hs-CRP: high sensitivity C-reactive protein.

Discussion

In this study, a clinical system to predict the effectiveness of *C sativus L* supplements in patients with allergic asthma was designed. The findings indicated that the model sufficiently estimated the level of alleviation in patients with asthma and the level of variation of asthma clinical, anti-inflammatory, and hematological factors.

Research [25] conducted on healthy individuals indicated that doses up to 400 mg per day of *C sativus L* were safe. In addition, a study [26] noted that a daily dosage of 100 mg of *C sativus L* in patients with metabolic syndrome reduced anti-HSP70 levels. Anti-HSP has been demonstrated to be a risk factor for asthma and has been correlated with asthma severity [8,26]. Another study [6] showed the beneficial effects of 100 mg *C sativus L* per day on clinical and immunologic factors as well as on symptom severity in asthma.

In a study [17], it was confirmed that it is possible to create a predictive model with machine learning algorithms that can outperform experts. Predictions can be in the form of patient classification, disease diagnosis, drug composition, and reactions to drugs. Prediction of treatment effects using machine learning have been used to develop clinical drug prediction systems [12,27]. To the best of our knowledge, our clinical prediction system is the first system that uses input features selection and classifies effectiveness of *C sativus L* supplement therapy.

The input features have an important role in the efficient implementation of prediction problems [28]. It is possible to use data mining methods to decrease the data dimensions, choose optimal features, and achieve better system precision [29]. In a study [15], partial least square regression was used to identify 9 out of 48 prognostic factors that were correlated to persistent asthma. Moreover, multilayer and probabilistic neural networks topologies were studied to find the best prediction accuracy [15]. In another study [29], data analysis was performed to select 13 effective factors to use in an ANN asthma diagnostic model [29]. We used a genetic algorithm feature selection model to find pertinent features. Genetic algorithms were recently developed to compare different feature selection methods and may be useful for feature selection when the problem has an exponential search space. The many advantages of genetic algorithms for feature selection are highlighted in the literature [30,31]. In contrast to that of other asthma diagnostic studies, the feature selection in our system is aimed to optimize the classification of asthma patients with supplement therapy into different groups in terms of the level of effect of *C sativus L* supplement.

ANN-based models have recently been used as a robust technique to classify patients with asthma, chronic obstructive pulmonary disease, or normal lung function based on measurement of lung condition and symptoms [7,32,33].

Analyses of the mean accuracies of the genetic algorithm–modified neural network predictor with feature selection included selecting effective risk factors and architecture components using an allergic asthma dataset for which the best MSE of 0.0004 was obtained. It was also clear that feature selection improved the accuracy of prediction for all asthma risk factors as well as the accuracy of the classification of patients in terms of level of alleviation.

Our study was able to show the importance prioritizing factors to predict the variation level of variation of factors in the patients with allergic asthma. Frequency of salbutamol inhaler use, frequency of waking due to asthma symptoms, and weight were the input features that were selected most often by the genetic algorithm to predict the level of variation in allergic asthma risk factors for the patients using *C sativus L* supplement therapy. Therefore, the initial value of those factors have an important role in predicting the level of alleviation of disease severity.

Several studies have listed the significant and nonsignificant effects of *C sativus L* supplements on the demographic, anthropometric, and clinical characteristics of asthma [6,8,9]. Hosseini et al [9] showed that *C sativus L* significantly affected anti-HSP and hs-CRP factor serum levels . Moreover, they showed that it increased the pulmonary volumes FEV_1, FVC, FEV_1/FVC, and $FEF_{25\%-75\%}$ [9]. They also showed that there was no significant effect on the eosinophil count [18]. The genetic algorithm–modified neural network system in this study accurately predicted the level of *C sativus L* effects for anti-HSP (96.5%), hs-CRP (98.9%), FEV_1 (98.1%), FVC (97.5%), FEV_1/FVC (97%), $FEF_{25\%-75\%}$ (96.7%), eosinophil (97.5%), and basophil (95.5%). Moreover, in general the patients were classified into 5 groups from low to high level of alleviation which reached accuracies of 99.9% in both training and testing experiments. FEV_1/FVC ratio, frequency of waking up due to asthma symptoms, hs-CRP, eosinophil, basophil, weight, and smoking history were selected as effective factors to estimate the level of alleviation of asthma in patients after the *C sativus L* supplement therapy.

The results also confirmed that it is possible to rely upon the prediction process described in this paper for the early prediction of level of variation of asthma factors. This study is the first to evaluate the classification accuracy of *C sativus L* supplement effect on the patients with asthma through feature selection. Our genetic algorithm–modified neural network can predict the effects of using *C sativus L* supplement on patients with asthma. This study indicated the importance of prioritizing each factor in predicting the effect of supplement therapy on allergic asthma.

By assessing risk, this method can be viewed as an important innovation to ensure that asthma is controlled and that serious complications are avoided. This study contributes to helping doctors to identify early on which factors will improve during treatment.

It would be interesting to evaluate the genetic algorithm–modified neural network on additional groups of patients with allergic asthma or other types of asthma. In future work, development of intelligent models including heuristic algorithms for feature selection with other machine learning techniques is recommended. Clinical prediction systems can also be applied to predict or to simulate the effectiveness of other medicines on other conditions, especially where different initial patient conditions may alter the course of treatment.

Acknowledgments

This research was financially supported by a grant from the Vice Chancellor for Research at Ahvaz Jundishapur University of Medical Sciences in Ahvaz, Iran (Research ID: 330096892). We would like to thanks Ahvaz Jundishapur University of Medical Sciences.

Conflicts of Interest

None declared.

References

1. To T, Stanojevic S, Moores G, Gershon AS, Bateman ED, Cruz AA, et al. Global asthma prevalence in adults: findings from the cross-sectional world health survey. BMC Public Health 2012 Mar 19;12(1):204 [FREE Full text] [doi: 10.1186/1471-2458-12-204] [Medline: 22429515]
2. Bateman ED, Hurd SS, Barnes PJ, Bousquet J, Drazen JM, FitzGerald M, et al. Global strategy for asthma management and prevention: GINA executive summary. Eur Respir J 2008 Jan 01;31(1):143-178 [FREE Full text] [doi: 10.1183/09031936.00138707] [Medline: 18166595]
3. Masoli M, Fabian D, Holt S, Beasley R, Global Initiative for Asthma (GINA) Program. The global burden of asthma: executive summary of the GINA Dissemination Committee report. Allergy 2004 May;59(5):469-478. [doi: 10.1111/j.1398-9995.2004.00526.x] [Medline: 15080825]
4. Pachter LM, Weller SC, Baer RD, de Alba Garcia JEG, Trotter RT, Glazer M, et al. Variation in asthma beliefs and practices among mainland Puerto Ricans, Mexican-Americans, Mexicans, and Guatemalans. J Asthma 2002 Apr 29;39(2):119-134. [doi: 10.1081/jas-120002193] [Medline: 11995676]

5. Boskabady MH, Farkhondeh T. Antiinflammatory, antioxidant, and immunomodulatory effects of Crocus sativus L. and its main constituents. Phytother Res 2016 Jul 21;30(7):1072-1094. [doi: 10.1002/ptr.5622] [Medline: 27098287]
6. Zilaee M, Hosseini SA, Jafarirad S, Abolnezhadian F, Cheraghian B, Namjoyan F, et al. An evaluation of the effects of saffron supplementation on the asthma clinical symptoms and asthma severity in patients with mild and moderate persistent allergic asthma: a double-blind, randomized placebo-controlled trial. Respir Res 2019 Mar 22;20(1):39 [FREE Full text] [doi: 10.1186/s12931-019-0998-x] [Medline: 30795753]
7. Do Q, Son TC, Chaudri J. Classification of asthma severity and medication using TensorFlow and multilevel databases. Procedia Computer Science 2017;113:344-351. [doi: 10.1016/j.procs.2017.08.343]
8. Yang M, Wu T, Cheng L, Wang F, Wei Q, Tanguay RM. Plasma antibodies against heat shock protein 70 correlate with the incidence and severity of asthma in a Chinese population. Respir Res 2005 Mar 14;6(1):18 [FREE Full text] [doi: 10.1186/1465-9921-6-18] [Medline: 15710045]
9. Hosseini SA, Zilaee M, Shoushtari MH, Ghasemi Dehcheshmeh M. An evaluation of the effect of saffron supplementation on the antibody titer to heat-shock protein (HSP) 70, hsCRP and spirometry test in patients with mild and moderate persistent allergic asthma: A triple-blind, randomized placebo-controlled trial. Respir Med 2018 Dec;145:28-34. [doi: 10.1016/j.rmed.2018.10.016] [Medline: 30509713]
10. Benjamins JW, Hendriks T, Knuuti J, Juarez-Orozco LE, van der Harst P. A primer in artificial intelligence in cardiovascular medicine. Neth Heart J 2019 Sep 20;27(9):392-402 [FREE Full text] [doi: 10.1007/s12471-019-1286-6] [Medline: 31111458]
11. Kagiyama N, Shrestha S, Farjo PD, Sengupta PP. Artificial iIntelligence: practical primer for clinical research in cardiovascular disease. JAHA 2019 Sep 03;8(17). [doi: 10.1161/jaha.119.012788]
12. Borisov N, Tkachev V, Muchnik I, Buzdin A. Individual drug treatment prediction in oncology based on machine learning using cell culture gene expression data. 2017 Oct 18 Presented at: ICCBB 2017: Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics; 2017; America.
13. Robin X, Turck N, Hainard A, Lisacek F, Sanchez J, Müller M. Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into in vitro diagnostics? Expert Rev Proteomics 2009 Dec 09;6(6):675-689. [doi: 10.1586/epr.09.83] [Medline: 19929612]
14. Arimoto R, Prasad M, Gifford EM. Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors. J Biomol Screen 2005 Apr;10(3):197-205. [doi: 10.1177/1087057104274091] [Medline: 15809315]
15. Chatzimichail E, Paraskakis E, Rigas A. Predicting asthma outcome using partial least square regression and artificial neural networks. Advances in Artificial Intelligence 2013;2013:1-7. [doi: 10.1155/2013/435321]
16. Demuth Howard. Neural network toolbox for use with MATLAB: User's guide. Neural network toolbox for use with MATLAB 1993 Dec 03;9.
17. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A 1982 Apr 15;79(8):2554-2558 [FREE Full text] [doi: 10.1073/pnas.79.8.2554] [Medline: 6953413]
18. Kareem Kamoona KR, Budayan C. Implementation of genetic algorithm integrated with the deep neural network for estimating at completion simulation. Advances in Civil Engineering 2019 May 02;2019:1-15. [doi: 10.1155/2019/7081073]
19. Holland JH. Adaptation in Natural and Artificial Systems. Cambridge: MIT Press; Apr 01, 1992:100.
20. Alharbi A, Alghahtani M. Using genetic algorithm and ELM neural networks for feature extraction and classification of type 2-diabetes mellitus. Applied Artificial Intelligence 2018 Dec 27;33(4):311-328. [doi: 10.1080/08839514.2018.1560545]
21. Chomatek L. Efficient Genetic Algorithm for Breast Cancer Diagnosis. In: Duraj A, editor. Information Technology in Biomedicine. Poland: Lodz University of Technology; Jan 01, 2019:64-76.
22. shokoufeh A, Hadi S, Alireza R, Saeid E. Feature selection using genetic algorithm for breast cancer diagnosisxperiment on three different datasets. Iranian journal of basic medical sciences 2016 May 01;19(5):476-482.
23. Shobhit K, Daya SS, Sapna S. Modified mean square error algorithm with reduced cost of training and simulation time for character recognition in backpropagation neural network. 2013 Jan 01 Presented at: Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA); 2013; Singapore p. 137-145. [doi: 10.1007/978-3-319-02931-3_17]
24. Amaury L, Vincent W, Michel V. Model selection with cross-validations and bootstraps?application to time series prediction with RBFN models. 2003 Jan 18 Presented at: International Conference on Artificial Neural Networks; 2003; Singapore p. 573-580. [doi: 10.1007/3-540-44989-2_68]
25. Modaghegh M, Shahabian M, Esmaeili H, Rajbai O, Hosseinzadeh H. Safety evaluation of saffron (Crocus sativus) tablets in healthy volunteers. Phytomedicine 2008 Dec;15(12):1032-1037. [doi: 10.1016/j.phymed.2008.06.003] [Medline: 18693099]
26. Shemshian M, Mousavi SH, Norouzy A, Kermani T, Moghiman T, Sadeghi A, et al. Saffron in metabolic syndrome: its effects on antibody titers to heat-shock proteins 27, 60, 65 and 70. Journal of Complementary and Integrative Medicine 2014 Feb 6;11(1). [doi: 10.1515/jcim-2013-0047] [Medline: 24501162]
27. Blumenschein GR, Saintigny P, Liu S, Kim ES, Tsao AS, Herbst RS, et al. Comprehensive biomarker analysis and final efficacy results of sorafenib in the BATTLE trial. Clinical Cancer Research 2013 Oct 28;19(24):6967-6975. [doi: 10.1158/1078-0432.ccr-12-1818]

28. Yang J, Nugroho AS, Yamauchi K, Yoshioka K, Zheng J, Wang K, et al. Efficacy of interferon treatment for chronic hepatitis C predicted by feature subset selection and support vector machine. J Med Syst 2007 Apr 21;31(2):117-123. [doi: 10.1007/s10916-006-9046-8] [Medline: 17489504]
29. Alizadeh B, Safdari R, Zolnoori M, Bashiri A. Developing an intelligent system for diagnosis of asthma based on artificial neural network. Acta Inform Med 2015 Aug;23(4):220-223 [FREE Full text] [doi: 10.5455/aim.2015.23.220-223] [Medline: 26483595]
30. Il-Seok Oh, Jin-Seon Lee, Byung-Ro Moon. Hybrid genetic algorithms for feature selection. IEEE Trans. Pattern Anal. Machine Intell 2004 Nov;26(11):1424-1437. [doi: 10.1109/tpami.2004.105]
31. Farzin H, Saadat V, Mehrnaz J. Quantitative structure-activity relationship studies of 4-imidazolyl-1, 4-dihydropyridines as calcium channel blockers. Iranian journal of basic medical sciences 2013 Aug 01;16(8):910-916. [doi: 10.5005/jp/books/11736_14]
32. Badnjevic A, Gurbeta L, Custovic E. An expert diagnostic system to automatically identify asthma and chronic obstructive pulmonary disease in clinical settings. Sci Rep 2018 Aug 03;8(1):11645 [FREE Full text] [doi: 10.1038/s41598-018-30116-2] [Medline: 30076356]
33. Hee HI, Balamurali B, Karunakaran A, Herremans D, Teoh OH, Lee KP, et al. Development of machine learning for asthmatic and healthy voluntary cough Sounds: a proof of concept study. Applied Sciences 2019 Jul 16;9(14):2833. [doi: 10.3390/app9142833]

Abbreviations

ANN: artificial neural network
BMI: body mass index
$FEF_{25\%-75\%}$: forced expiratory flow, midexpiratory phase
FEV_1: forced expiratory volume in the first second of expiration
FVC: forced vital capacity
hs-CRP: high-sensitivity C-reactive protein
HSP: heat shock protein
MSE: mean squared error
WHO: World Health Organization

Edited by G Eysenbach; submitted 21.12.19; peer-reviewed by SMH Mousavi Jazayeri, A Pazahr, M Ghayour-mobarhan; comments to author 17.01.20; revised version received 22.02.20; accepted 26.02.20; published 06.07.20.

Please cite as:
Hosseini SA, Jamshidnezhad A, Zilaee M, Fouladi Dehaghi B, Mohammadi A, Hosseini SM
Neural Network–Based Clinical Prediction System for Identifying the Clinical Effects of Saffron (Crocus sativus L) Supplement Therapy on Allergic Asthma: Model Evaluation Study
JMIR Med Inform 2020;8(7):e17580
URL: https://medinform.jmir.org/2020/7/e17580
doi:10.2196/17580
PMID:32628613

Original Paper

Barriers and Facilitators to Implementation of Medication Decision Support Systems in Electronic Medical Records: Mixed Methods Approach Based on Structural Equation Modeling and Qualitative Analysis

Se Young Jung[1,2*], MD, MPH; Hee Hwang[1,3*], MD, PhD; Keehyuck Lee[1,2], MD, MBA; Ho-Young Lee[1,4], MD, PhD; Eunhye Kim[1], RN; Miyoung Kim[1], RN; In Young Cho[2], MD

[1]Office of eHealth Research and Businesses, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[2]Department of Family Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[3]Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[4]Department of Nuclear Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Keehyuck Lee, MD, MBA
Office of eHealth Research and Businesses
Seoul National University Bundang Hospital
Dolma-ro 172, Bundang-gu
Seongnam, 13605
Republic of Korea
Phone: 82 317878992
Email: chrisruga@naver.com

Abstract

Background: Adverse drug events (ADEs) resulting from medication error are some of the most common causes of iatrogenic injuries in hospitals. With the appropriate use of medication, ADEs can be prevented and ameliorated. Efforts to reduce medication errors and prevent ADEs have been made by implementing a medication decision support system (MDSS) in electronic health records (EHRs). However, physicians tend to override most MDSS alerts.

Objective: In order to improve MDSS functionality, we must understand what factors users consider essential for the successful implementation of an MDSS into their clinical setting. This study followed the implementation process for an MDSS within a comprehensive EHR system and analyzed the relevant barriers and facilitators.

Methods: A mixed research methodology was adopted. Data from a structured survey and 15 in-depth interviews were integrated. Structural equation modeling was conducted for quantitative analysis of factors related to user adoption of MDSS. Qualitative analysis based on semistructured interviews with physicians was conducted to collect various opinions on MDSS implementation.

Results: Quantitative analysis revealed that physicians' expectations regarding ease of use and performance improvement are crucial. Qualitative analysis identified four significant barriers to MDSS implementation: alert fatigue, lack of accuracy, poor user interface design, and lack of customizability.

Conclusions: This study revealed barriers and facilitators to the implementation of MDSS. The findings can be applied to upgrade MDSS in the future.

(JMIR Med Inform 2020;8(7):e18758) doi:10.2196/18758

KEYWORDS

clinical decision support system; electronic health record; medication safety; Computerized Provider Order Entry (CPOE)

Introduction

Background

In 2009, based on evidence that electronic health records (EHR) can improve healthcare quality, the US government enacted the Health Information Technology for Economic and Clinical Health (HITECH) Act [1]. Over the past decade, the healthcare industry has experienced a tremendous digital revolution initiated by the government's efforts to implement EHRs [2,3]. As of 2017, over 90% of general medical and surgical hospitals in the US use certified EHR systems, thus generating an enormous amount of electronic medical information daily [2,4]. Analysis of big data gathered from EHRs can generate real-time evidence that helps end-users take better care of their patients [5]. Clinical decision support systems (CDSS) are a typical example of value provided to EHR users [6]. Such systems intervene in real-time to help users make appropriate decisions based on up-to-date information from EHRs. Medication decision support systems (MDSS), a well-known and frequently used type of CDSS, reduce adverse drug events (ADE), some of the most common causes of iatrogenic injuries in hospitals [7]. ADEs are generally defined as anticipated or unanticipated side effects resulting primarily from medication errors, attributable to human errors. The most common types of medication errors include the use of contraindicated drugs and overdosing [8]. An MDSS checks for problems based on CDSS data and alerts users in advance of potentially preventable errors. However, despite the high adoption rate of EHRs in the US, ADEs are still a significant problem [9]. The situation is similar in South Korea. ADEs have not been reduced dramatically in South Korea, although the adoption rate of EHRs is around 90% as of 2017 [10,11].

Prior Research

Efforts have been made to reduce medication errors to minimize the frequency of ADEs. Previous studies have demonstrated that implementing an MDSS in the EHR system improves patient care and overall outcomes by reducing medication errors [12-19]. However, repeated false alerts from MDSSs can decrease healthcare professionals' productivity by interrupting their workflow [14,20-22]. Furthermore, if doctors are frequently interrupted by false alerts, they are less likely to adopt MDSS recommendations [23].

Research has shown that physicians override about 90% of drug allergy and high-severity drug interaction warning notifications [20,21,24,25]. Two methods could be adopted to improve MDSS. One is to enhance the precision of the MDSS algorithms to reduce unhelpful notifications. Machine learning techniques are being widely considered to provide personalized, accurate notifications [15]. The other method is to support users by understanding the factors associated with MDSS feasibility and usability, which requires an understanding of what factors users consider important in clinical settings. To date, many studies have explored the effectiveness of MDSS, but only a few have analyzed their feasibility and usability.

Aim

This study aimed to analyze factors related to the adoption of MDSS. A mixed-methods research approach was taken to both quantitatively measure factors necessary for the successful implementation of MDSS and to qualitatively gather and reflect on the opinions of end users. The study encompassed the entire process of implementing an MDSS into a comprehensive EHR system and analyzed the relevant barriers and facilitators. Based on the results, some ideas are suggested to support users and upgrade MDSS, resolving issues already well established by previous studies.

Methods

Design

A mixed research methodology was adopted [26]. Data from a structured survey and 15 in-depth interviews were integrated, and structural equation modeling (SEM) was conducted to yield a quantitative analysis of the factors related to user adoption of MDSS. A qualitative analysis based on semistructured interviews with physicians also collected various opinions about MDSS implementation. To objectively report results, the qualitative analysis followed the Consolidated Criteria for Reporting Qualitative (COREQ) Research Guidelines [27].

Setting and Participants

The study was conducted at Seoul National University Bundang Hospital (SNUBH), where the comprehensive, privately developed BESTCare electronic medical record (EMR) has been in use since 2003. The system has been accredited three times as a Health Information Management Systems and Society Analytics EMR Adoption Model Stage 7 since 2010. BESTCare implements a proprietary MDSS concurrently with a prescription drug monitoring program run by the South Korean government; thus, BESTCare users are already familiar with MDSS.

A taskforce team of 12 attending physicians, two pharmacists, three nurses, and three engineers was formed to improve medication safety. The team decided to introduce a third-party MDSS to BESTCare. In November 2016, the team analyzed MDSS previously released in the market and decided to implement the Medi-Span solution. The task force analyzed mapping codes, mapping contents, and filters, and designed the user interface and overall system architecture (Table 1).

The taskforce team designed the overall system architecture of the MDSS (Figures 1 and 2) and designed alert screens to display messages efficiently (Figure 3).

After implementing the MDSS in April 2017, we conducted a structured survey of physicians between May 2017 and October 2017 and employed SEM to analyze the factors facilitating successful implementation. Focused group interviews were also conducted to collect direct opinions from end-users. For the qualitative analysis, study participants were selected through purposive sampling [28], aiming to include participants with in-depth knowledge of the work process involving the EHR system and MDSS. Textbox 1 presents the items included in the semistructured interview questionnaire.

Table 1. Basic considerations for integration of a commercialized MDSS into the EMR system.

Mapping	What standard codes should be used to interface Medi-Span with BESTCare?
Contents	What functions should be implemented to improve medication safety? • Drug interactions • Drug allergies • Drug disease contraindications • Duplicate therapy • Dose screening and drug order • Route contraindications • Pregnancy/lactation contraindications • Gender/age contraindications
Filters	What filters should we integrate? How can we control users' authorization to override MDSS alerts?
Alerts	How can we show alerts efficiently?
User interface/experience	How can we improve user experience and user interface designs?

Figure 1. MDSS system architecture and configuration. EMR: electronic medical record; HIS: hospital information system.

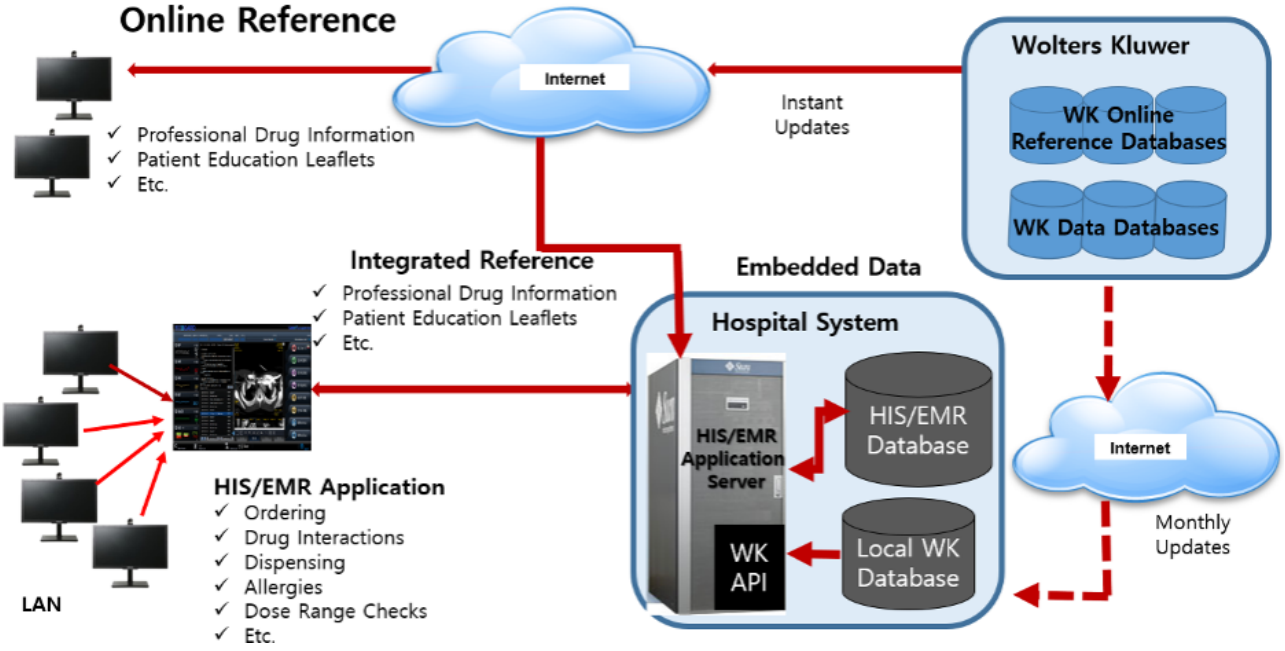


Figure 2. MDSS function list. API: application programming interface; GPI: generic product identifier.

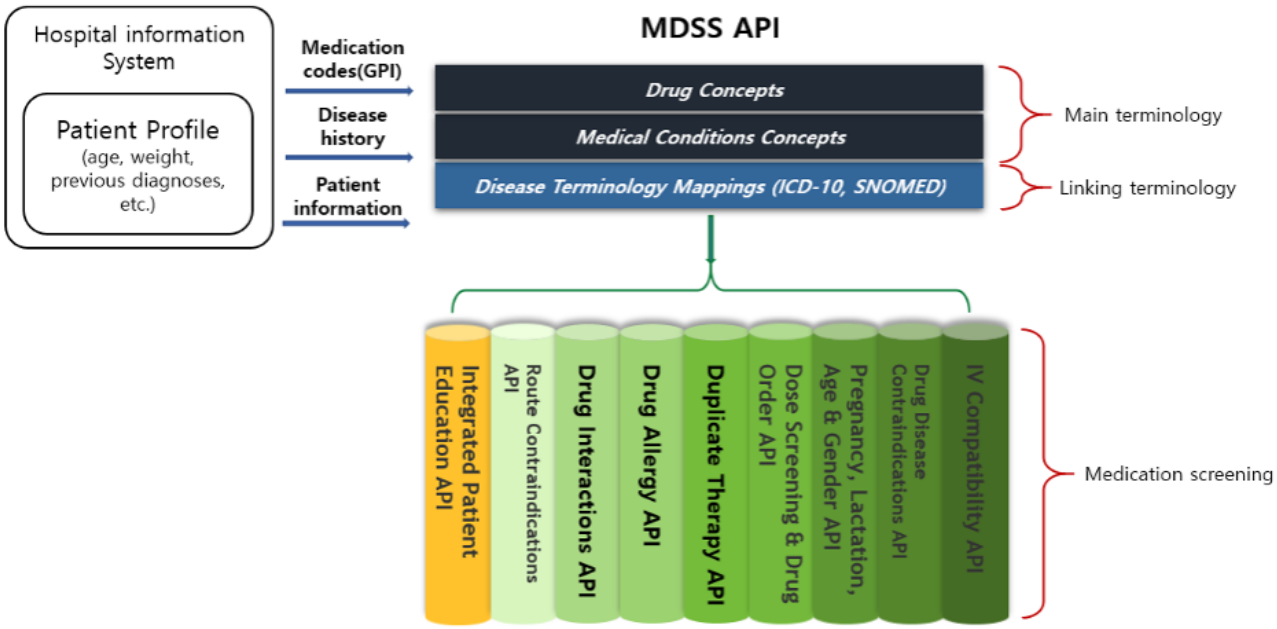
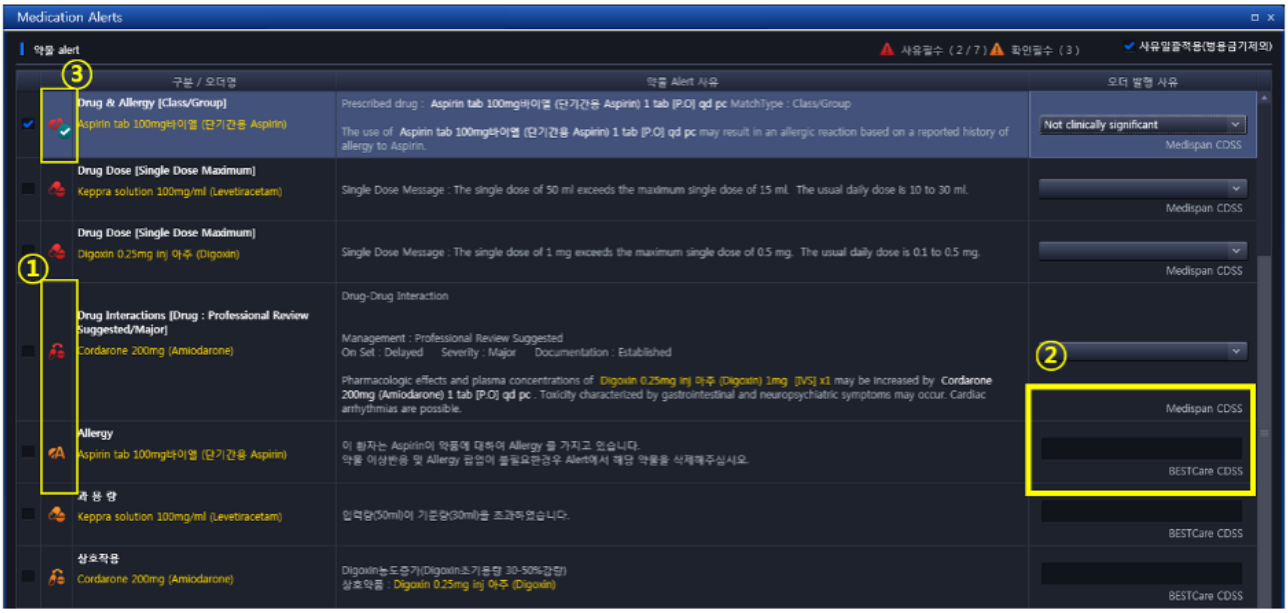


Figure 3. Screenshot of the MDSS user interface. 1) Override requirements: red alerts indicate that users must view an alert message and select a reason for overriding it, whereas orange alerts indicate that users must confirm the alert. 2) Origin of alerts: BESTCare MDSS, Medi-span CDSS, and South Korean national prescription drug monitoring program. 3) Classification of alerts using icons, allowing users to see the notification easily.



Textbox 1. Semistructured interview questionnaire items.

Questions:

- What was your first impression of the MDSS implemented in this hospital?
- Did you have experience with other MDSSs before?
- How long did it take for you to get used to the MDSS?
- Were there any barriers to implementation of the MDSS?
- What do you think would help to better implement the MDSS?
- How is the navigation when using the MDSS?
- Are there any problems with the MDSS that need to be resolved?
- Do you think the MDSS is customized well for the EHR workflow? Is there anything missing?
- What features or functions do you want to add to the MDSS?
- Do you have any recommendations for the MDSS to improve your work experience?
- Is there anything else you want to mention regarding the MDSS?

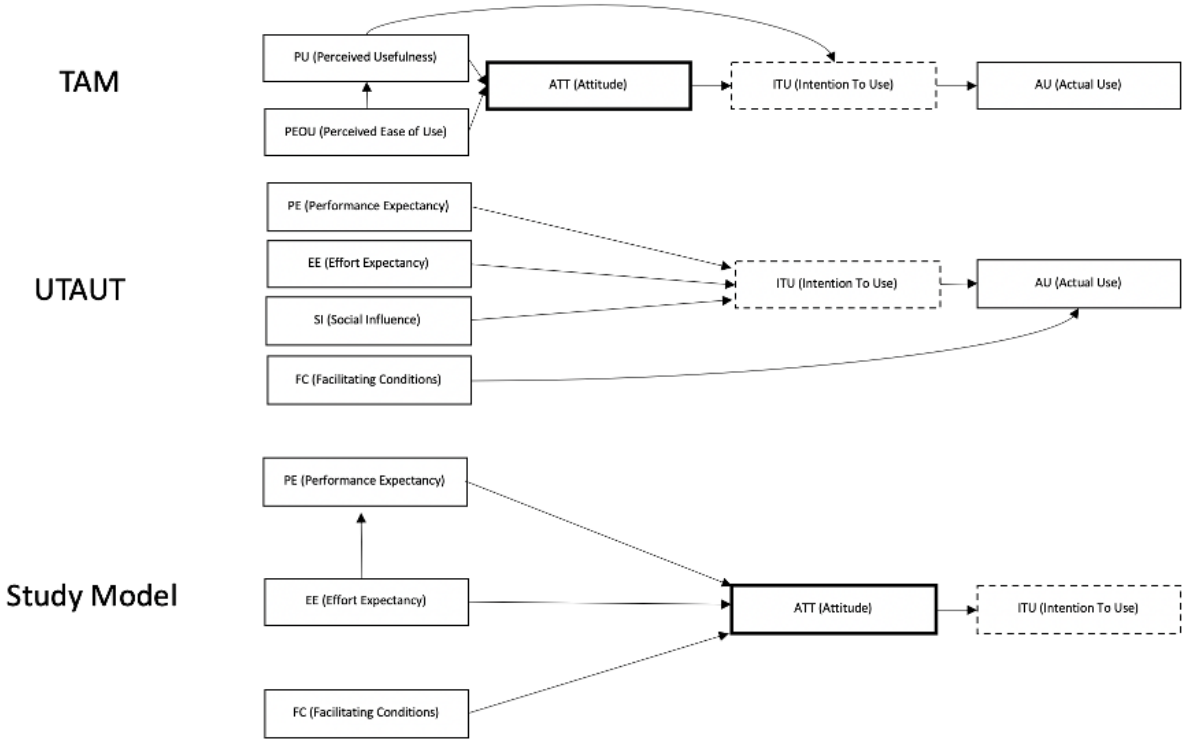
Data Collection

MYK, a registered nurse, conducted the survey and face-to-face semistructured interviews. IYC, a medical doctor, also led the interviews and took notes. Both interviewers received training on qualitative interviews. The interviews lasted 20 to 60 minutes and were recorded in a closed office or conference room. Nobody was present besides the participants and researchers. During the sessions, MYK followed the semistructured interview questionnaire covering topics related to the implementation of the MDSS (Textbox 1). The researchers followed interview guidelines based on previous research and approved by members of the eHealth research team at SNUBH.

Data Analysis

For SEM, the survey adopted the technology acceptance model (TAM) and the unified theory of acceptance and use of technology (UTAUT), both of which have been widely adopted to analyze user willingness to accept new technologies [15,29-32]. The models were modified to create a structural equation model optimized for this study Figure 4. Performance expectancy, effort expectancy, and facilitating conditions were expected to have a positive influence on attitude, and attitude was expected to have a positive influence on intention to use. The TAM includes two variables impacting behavioral intentions to use, and the UTAUT includes three behavioral variables and one variable that influences actual use, all of which influence the overall process. Theoretically, facilitating conditions should influence actual use. However, based on previous studies, we hypothesized that facilitating conditions would instead moderate intention to use due to difficulties in measuring actual use. Social influence from senior colleagues was omitted in order to simplify the model, as the study's focus was on factors related only to user expectations and support from the hospital.

Figure 4. The analytical model used in this study, modified from the technology acceptance model (TAM) and the unified theory of acceptance and use of technology (UTAUT).



Ethics

The research protocol was approved by the Institutional Review Board of Human Research of Seoul National University Bundang Hospital (Protocol No. B-1709-420-303).

Results

Participant Demographics

Table 2 presents the demographic characteristics of all SEM survey respondents. Of the 80 professionals invited to take the survey, 61 responded. Most were residents who use the EMR and MDSS more actively than any other position on the hospital staff.

For qualitative analysis, 15 people out of 80 participants were interviewed. Table 3 presents the interviewees' demographic characteristics.

A reliability test was performed to confirm the consistency of the survey items for SEM analysis. Cronbach α exceeded .8 for all variables except facilitating conditions. Thus, the survey items were confirmed to be consistent and reliable (Table 4).

Table 2. Demographic characteristics of SEM survey respondents.

Categories/items	Number	Percentage
Gender		
Male	22	36.07
Female	39	63.93
Age (years)		
20-29	13	21.31
30-39	45	73.77
40-49	2	3.28
50 and above	1	1.64
Department		
Internal/family medicine	42	68.85
Pediatrics	8	13.11
Surgery	5	8.20
Other	6	9.84
Length of service (years)		
<1	24	39.34
1-3	30	49.18
3-5	4	6.56
5-10	0	0
>10	3	4.92
Position		
Professor	2	3.28
Fellow	4	6.56
Resident	55	90.16

Table 3. Demographic characteristics of focused interview participants.

Categories/items	Number	Percentage
Gender		
Male	11	73
Female	4	27
Age (years)		
20-29	2	7
30-39	12	80
40-49	1	13
Department		
Internal/family medicine	13	86
Pediatrics	1	7
Surgery	1	7
Length of service		
<1	1	7
1-3	12	79
3-5	1	7
5-10	1	7
Position		
Professor	3	20
Resident	12	80

Table 4. Reliability analysis.

Construct	Number of items	Cronbach α
Performance expectancy	3	.82
Effort expectancy	3	.90
Attitude	2	.86
Facilitating conditions	3	.70
Intention to use	3	.95

Quantitative Analysis

Table 5 shows the total number of red and orange alerts presented by the MDSS each month from April 2017 to March 2018. A total of 185,441 red alerts (65.82%) were overridden.

Table 6 presents the usability test results. The overall mean score was 3.38. Study participants generally agreed with all statements except “I feel confident using Medi-Span,” which resulted in a positive score (above 3) but was not statistically significant.

Figure 5 presents the overall results of the SEM analysis. For the research model, χ^2 was 93.51 (df=60, P<.01), the Tucker-Lewis index (TLI) was 0.971, the comparative fit index (CFI) was 0.987, and the root mean square error of approximation (RMSEA) was 0.067. Because TLI and CFI exceeded 0.9, RMSEA was below 0.1, and the P value of the model was statistically significant, the model was confirmed to be appropriate for analyzing end-user intentions to use the MDSS. The associations between latent variables were positive and statistically significant, confirming the influence of performance expectancy on attitude, effort expectancy on performance expectancy, and attitude on the intention to use.

Table 5. Number of red and orange alerts presented by the MDSS each month.

Month	Red alerts (*N*)	Orange alerts (*N*)
April 2017	789	3805
May 2017	3979	29,037
June 2017	25,903	94,023
July 2017	23,868	73,986
August 2017	27,468	74,070
September 2017	24,401	69,403
October 2017	22,729	62,991
November 2017	25,536	72,996
December 2017	28,445	76,197
January 2018	27,728	82,671
February 2018	37,624	68,926
March 2018	3250	78,162
Total	1,126,724	281,720

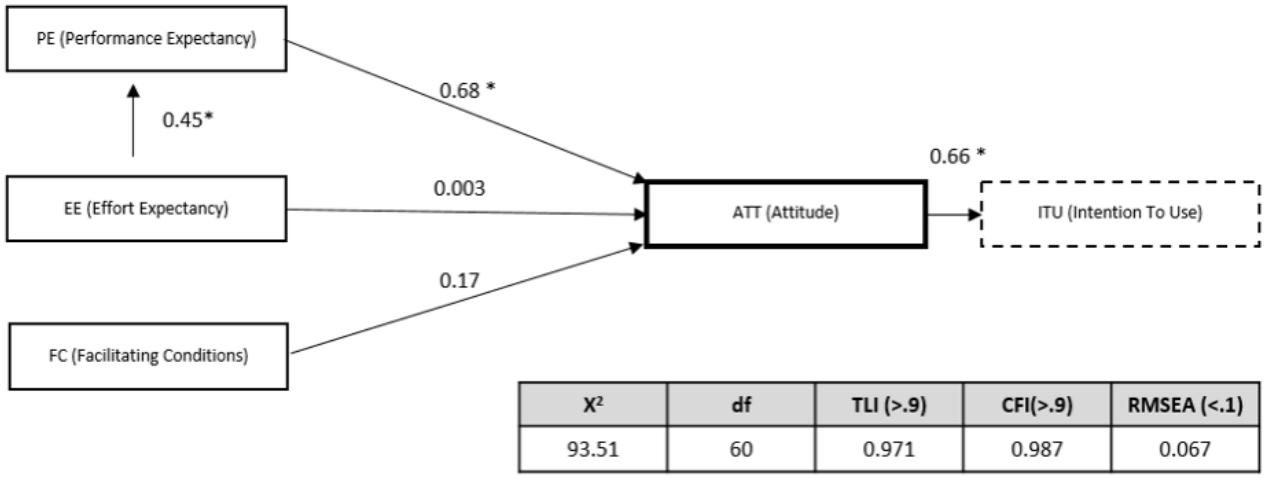
Table 6. Usability test results.

Items[a]	Mean (95% CI)
I feel like I use Medi-Span frequently.	3.43 (3.23, 3.62)
Medi-Span is unnecessarily complicated to use.[a]	2.74 (3.08, 3.45)
Medi-Span is easy to use.	3.46 (3.31, 3.61)
I need technical support to use Medi-Span.[a]	2.70 (2.47, 2.93)
Medi-Span integrates various functions well.	3.41 (3.26, 3.56)
Medi-Span is not consistent in terms of usability.[a]	2.48 (2.31, 2.63)
I think most people learn how to use Medi-Span quickly.	3.51 (3.34, 3.68)
It is bothersome to use Medi-Span.[a]	2.59 (2.31, 2.67)
I feel confident using Medi-Span.	2.93 (2.78, 3.08)
It takes a long time to get used to Medi-Span.[b]	2.61 (2.45, 2.77)
Total score	3.38 (3.27, 3.48)

[a]Each item was rated on a 5-point Likert scale, with a score of 3 or higher indicating agreement with the statement (for questions with negative wording, a score of 3 or below indicated a positive response).

[b]Questions with negative wording were reverse-scored to calculate the mean total score.

Figure 5. Results of the SEM model. *P<.001.



X^2	df	TLI (>.9)	CFI(>.9)	RMSEA (<.1)
93.51	60	0.971	0.987	0.067

Qualitative Analysis

Alert Fatigue

Alert fatigue is a well-known problem associated with MDSS [20]. Participants in this study also mentioned alert fatigue several times.

> *It is inconvenient because there are many alerts for drugs commonly used in hematology-oncology, and it is not possible to dismiss the extreme caution alert, which frequently appears in older patients.*

Lack of Accuracy

Accuracy is an essential factor affecting users' trust in an MDSS. If false alerts pop up repeatedly, users will tire of the interruption and may fail to take heed when a valid alert is given. Accuracy is also closely related to alert fatigue because poor accuracy results in a higher number of unnecessary alerts. This study's participants also mentioned accuracy frequently.

> *For example, when co-prescribing morphine and clopidogrel, the same message about drug-drug interaction occurs several times, and the alert override has to be selected several times.*

Poor User Interface Design

South Korean medical staff are accustomed to using English at work. However, poor user interface design can display too much English information on one screen, making it difficult for professionals to see every message in a busy hospital setting. Particularly in emergent situations, poor user interface design presenting excessive and unnecessary English information can be problematic. Therefore, it is crucial to design a user interface that provides essential messages only. In this study, two participants mentioned that the context of English information was difficult to understand quickly.

> *It's hard to understand the alert messages because they're in English and include a lot of content.*

Lack of Customizability

Participants highlighted the need for functionality in the MDSS to customize types of alerts according to user preference. To reduce the rate of overrides and alert fatigue, an MDSS must be easily customizable.

> *It would be nice to have the ability to set specific drugs and specific doses as a basis for alerts for each department or doctor.*

Discussion

Principal Findings

This study employed a mixed-methods approach to analyze barriers and facilitators to the implementation of MDSS. Barriers were identified based on the results of SEM and qualitative analysis, and facilitators were identified based on SEM.

The quantitative analysis found an average usability rating of 3.38 out of 5, indicating acceptable usability of the MDSS. SEM analysis revealed that effort expectancy had a positive effect on performance expectancy, performance expectancy had a positive effect on attitude, and attitude had a positive effect on the intention to use. Thus, user expectations regarding ease of use may not directly affect their attitude. If users can utilize the system easily, they expect it to result in performance improvement, which in turn affects their attitude toward and intentions to use the system.

The qualitative analysis identified four significant barriers to implementation of an MDSS: alert fatigue, lack of accuracy, poor user interface design, and lack of customizability. To our knowledge, this is the first study to analyze barriers, facilitators, and usability of MDSS implementation based on a mixed-methods approach.

Barriers

Alert Fatigue

A previous study revealed that medication safety alert fatigue could be reduced through interaction design and clinical role tailoring [33]. The results of our SEM analysis revealed that effort expectancy (ie, user expectations regarding ease of use) affects performance expectancy, which in turn affects intentions to use the system. If the problem of frequent alert fatigue is neglected, the usability of the MDSS will suffer, which will affect user performance expectancy and foster a negative attitude

toward the use of the MDSS. In particular, previous studies have shown that the busy working environment of interns or residents can aggravate the adverse effects of alert fatigue [22].

Poor User Interface Design

Users clearly want to improve their work performance by using the MDS system. If they can utilize the MDSS to its fullest extent, they can expect to increase their work efficiency and performance. User interface design and experience are crucial to facilitate full utilization. If the system is difficult to use (ie, navigation is unintuitive), doctors will tend to dismiss or ignore messages from the MDSS. Previous studies have shown that user interface design is an essential factor in the successful implementation of a CDSS [34-36]. BESTCare has an integrated interface design, allowing users to easily and intuitively predict the next necessary action. The same interface design was adopted and upgraded to implement a third-party MDSS, helping users to quickly grasp the information presented by MDSS and move on to the next required action. However, the qualitative analysis found that redundant information hindered the user interface of the MDSS. Thus, the volume and layout of the displayed information must be considered in addition to the screen design.

Lack of Accuracy

A previous study found that 52.6% of MDSS alerts in outpatient clinics were overridden, 53% of which were appropriate [21]. Another study showed that 73.3% of patient allergy, duplicate drug, and drug interaction alerts were overridden in an inpatient clinic, only about 60% of which were appropriate [20]. Even if only 40% of overrides are considered inappropriate, this can significantly increase the risk of medication errors and potentially leading to ADEs. The override rate was 65.82% in the MDSS evaluated in this study, similar to previous results. Accuracy is related to performance expectancy. If accuracy remains consistently low, users will begin to lose hope of improving performance, resulting in a negative cycle of higher alert overrides.

Poor Customizability

BESTCare, the EHR integrated with the Medi-Span based MDSS used in this study, has an alert-related authority control function that meets the standards of the EHR certification program run by the Office of the National Coordinator for Health Information Technology. The professionals who participated in this study's in-depth interview were dissatisfied with this integrated management system and wanted the ability to customize and adjust the alerts they received. MDS systems are usually introduced to prevent ADEs. Therefore, a centralized, integrated management system is necessary for consistency and stability. However, end-user satisfaction will increase if they can adjust the level of alerts provided without sacrificing this overall stability.

Facilitators

Effort Expectancy

This study's quantitative analysis found that doctors generally considered the MDSS easy to use. Before development, both hospital staff and developers expressed concern about integrating the new Medi-Span MDSS on top of the two MDSS already integrated into the EHR system because the integration of three MDSS within the EHR could result in an excessive number of alerts. However, dedicated trial and error by the taskforce team ensured the usability of the system. As revealed in the in-depth interview, the users' wishes regarding MDSS usability can never be fully satisfied. However, the taskforce team's activities to improve usability acted as important facilitators for the successful introduction of the MDSS.

Performance Expectancy

According to this study's SEM analysis, end-user effort expectancy had a positive effect on their expectations of performance improvement. MDSS platforms must provide users with feedback on their actions in response to alerts and performance improvement outcomes in order to reduce overrides for valid alerts. Previous studies have noted that gaining user trust is crucial for the proper implementation and maintenance of a new system [37-39]. Clinical indicators regarding performance improvement and regular result reporting may be an excellent way to promote the use of the MDSS and gain trust. For example, public disclosure of antibiotic use rates effectively lowered the use of antibiotics for upper respiratory infections in South Korea [40]. Likewise, public disclosure about ADEs prevented by using the MDSS may help reduce alert override rates. Another option is to create a method by which users can provide feedback on false alerts. If doctors can provide feedback about false alerts instead of using the MDSS passively, the system can be dynamically upgraded to gain trust.

Limitations and Future Research

This study's main limitation is that the system was implemented in only one hospital. External validation in other hospitals is needed to help generalize the study results. Nevertheless, this research demonstrated the effects of interaction between user expectations regarding ease of use and performance improvement on their attitude toward using an MDSS, which can inform the practices of system designers and policymakers in charge of MDSS development.

Conclusion

This study revealed barriers and facilitators to the implementation of MDSS. The study's findings can be used as a reference to upgrade MDSSs effectively. Further studies are needed to evaluate specific ways to gain MDSS users' trust.

Conflicts of Interest

There are no conflicts of interest that could influence the findings of this research.

References

1. Blumenthal D. Stimulating the Adoption of Health Information Technology. N Engl J Med 2009 Apr 09;360(15):1477-1479. [doi: 10.1056/nejmp0901592]
2. Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. Health Affairs 2017 Aug;36(8):1416-1422. [doi: 10.1377/hlthaff.2016.1651]
3. GOLD M, McLAUGHLIN C. Assessing HITECH Implementation and Lessons: 5 Years Later. The Milbank Quarterly 2016 Sep 13;94(3):654-687. [doi: 10.1111/1468-0009.12214]
4. Percent of Specialty Hospitals that Possess Certified Health IT. The Office of the National Coordinator for Health Information Technology. URL: https://dashboard.healthit.gov/quickstats/pages/specialty-hospital-ehr-adoption.php [accessed 2020-02-13]
5. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014 Feb 7;2(1). [doi: 10.1186/2047-2501-2-3]
6. Wasylewicz A, Scheepers-Hoeks A. Clinical Decision Support Systems. Kubben P, Dumontier M, Dekker A. eds. Fundamentals of Clinical Data Science. Cham (CH): : Springer 2018 Dec 22:2018. [doi: 10.1007/978-3-319-99713-1_11] [Medline: 31314237]
7. Kohn L, Corrigan J, Donaldson M. To err is human: building a safer health system. National academy press Washington, DC 2000 Jan 01:2000. [doi: 10.17226/9728] [Medline: 25077248]
8. Zafar A, Hickner J, Pace W. An adverse drug event and medication error reporting system for ambulatory care (MEADERS). 2008 Nov 06 Presented at: AMIA Annu Symp Proc ;?43; 2008; Washinton p. 839.
9. Harris Y, Hu DJ, Lee C, Mistry M, York A, Johnson TK. Advancing Medication Safety: Establishing a National Action Plan for Adverse Drug Event Prevention. Jt Comm J Qual Patient Saf 2015 Aug;41(8):351-360. [doi: 10.1016/s1553-7250(15)41046-3] [Medline: 26215524]
10. Park Y, Han D. Current Status of Electronic Medical Record Systems in Hospitals and Clinics in Korea. Healthc Inform Res 2017 Jul;23(3):189-198 [FREE Full text] [doi: 10.4258/hir.2017.23.3.189] [Medline: 28875054]
11. Cho M, Kang DY, Kang H. Adverse drug reactions. J Korean Med Assoc 2019;62(9):472. [doi: 10.5124/jkma.2019.62.9.472]
12. Jia P, Zhang L, Chen J, Zhao P, Zhang M. The Effects of Clinical Decision Support Systems on Medication Safety: An Overview. PLoS ONE 2016 Dec 15;11(12):e0167683. [doi: 10.1371/journal.pone.0167683]
13. Salmasian H, Tran TH, Chase HS, Friedman C. Medication-indication knowledge bases: a systematic review and critical appraisal. J Am Med Inform Assoc 2015 Sep 02:ocv129. [doi: 10.1093/jamia/ocv129]
14. Payne T, Hines L, Chan R, Hartman S, Kapusnik-Uner J, Russ AL, et al. Recommendations to improve the usability of drug-drug interaction clinical decision support alerts. J Am Med Inform Assoc 2015 Nov;22(6):1243-1250. [doi: 10.1093/jamia/ocv011] [Medline: 25829460]
15. Esmaeilzadeh P, Sambasivan M, Kumar N, Nezakati H. Adoption of clinical decision support systems in a developing country: Antecedents and outcomes of physician's threat to perceived professional autonomy. Int J Med Inform 2015 Aug;84(8):548-560. [doi: 10.1016/j.ijmedinf.2015.03.007] [Medline: 25920928]
16. Felkey BG, Fox BI. Consider the benefits of a fully integrated medication use process. Hosp Pharm 2014 Jan;49(1):101-102 [FREE Full text] [doi: 10.1310/hpj4901-101] [Medline: 24421567]
17. Tawadrous D, Shariff SZ, Haynes RB, Iansavichus AV, Jain AK, Garg AX. Use of clinical decision support systems for kidney-related drug prescribing: a systematic review. Am J Kidney Dis 2011 Dec;58(6):903-914. [doi: 10.1053/j.ajkd.2011.07.022] [Medline: 21944664]
18. Nieuwlaat R, Connolly SJ, Mackay JA, Weise-Kelly L, Navarro T, Wilczynski NL, CCDSS Systematic Review Team. Computerized clinical decision support systems for therapeutic drug monitoring and dosing: a decision-maker-researcher partnership systematic review. Implement Sci 2011 Aug 03;6(1):90 [FREE Full text] [doi: 10.1186/1748-5908-6-90] [Medline: 21824384]
19. Hemens BJ, Holbrook A, Tonkin M, Mackay JA, Weise-Kelly L, Navarro T, CCDSS Systematic Review Team. Computerized clinical decision support systems for drug prescribing and management: a decision-maker-researcher partnership systematic review. Implement Sci 2011 Aug 03;6(1):89 [FREE Full text] [doi: 10.1186/1748-5908-6-89] [Medline: 21824383]
20. Nanji K, Seger D, Slight S, Amato MG, Beeler PE, Her QL, et al. Medication-related clinical decision support alert overrides in inpatients. J Am Med Inform Assoc 2018 May 01;25(5):476-481. [doi: 10.1093/jamia/ocx115] [Medline: 29092059]
21. Nanji KC, Slight SP, Seger DL, Cho I, Fiskio JM, Redden LM, et al. Overrides of medication-related clinical decision support alerts in outpatients. J Am Med Inform Assoc 2014 May 01;21(3):487-491 [FREE Full text] [doi: 10.1136/amiajnl-2013-001813] [Medline: 24166725]
22. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017 Apr 10;17(1):36 [FREE Full text] [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
23. Tsai C, Wang S, Hsu M, Li Y. Do false positive alerts in naïve clinical decision support system lead to false adoption by physicians? A randomized controlled trial. Comput Methods Programs Biomed 2016 Aug;132:83-91. [doi: 10.1016/j.cmpb.2016.04.011] [Medline: 27282230]
24. Isaac T, Weissman JS, Davis RB, Massagli M, Cyrulik A, Sands DZ, et al. Overrides of medication alerts in ambulatory care. Arch Intern Med 2009 Feb 09;169(3):305-311. [doi: 10.1001/archinternmed.2008.551] [Medline: 19204222]

25. Weingart SN, Toth M, Sands DZ, Aronson MD, Davis RB, Phillips RS. Physicians' decisions to override computerized drug alerts in primary care. Arch Intern Med 2003 Nov 24;163(21):2625-2631. [doi: 10.1001/archinte.163.21.2625] [Medline: 14638563]
26. Slootman M. A Mixed-Methods Approach. In: Slootman M. ed. Ethnic Identity, Social Mobility and the Role of Soulmates. Cham: : Springer International Publishing 2018. 41?. Switzerland: Springer, Cham; Sep 22, 2018:57.
27. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec 16;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
28. Robinson OC. Sampling in Interview-Based Qualitative Research: A Theoretical and Practical Guide. Qualitative Research in Psychology 2013 Nov 18;11(1):25-41. [doi: 10.1080/14780887.2013.801543]
29. Davis FD. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
30. Venkatesh, Morris, Davis, Davis. User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
31. Holden RJ, Karsh B. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
32. Gagnon M, Desmartis M, Labrecque M, Car J, Pagliari C, Pluye P, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb 30;36(1):241-277 [FREE Full text] [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
33. Hussain M, Reynolds T, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. J Am Med Inform Assoc 2019 Oct 01;26(10):1141-1149. [doi: 10.1093/jamia/ocz095] [Medline: 31206159]
34. Campion TR, Waitman LR, Lorenzi NM, May AK, Gadd CS. Barriers and facilitators to the use of computer-based intensive insulin therapy. Int J Med Inform 2011 Dec;80(12):863-871 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.10.003] [Medline: 22019280]
35. Kanstrup A, Christiansen M, Nøhr C. Four principles for user interface design of computerised clinical decision support systems. Stud Health Technol Inform 2011;166:65-73. [Medline: 21685612]
36. Yuan MJ, Finley GM, Long J, Mills C, Johnson RK. Evaluation of user interface and workflow design of a bedside nursing clinical decision support system. Interact J Med Res 2013 Jan 31;2(1):e4 [FREE Full text] [doi: 10.2196/ijmr.2402] [Medline: 23612350]
37. Koskela T, Sandström S, Mäkinen J, Liira H. User perspectives on an electronic decision-support tool performing comprehensive medication reviews - a focus group study with physicians and nurses. BMC Med Inform Decis Mak 2016 Jan 22;16(1):6 [FREE Full text] [doi: 10.1186/s12911-016-0245-z] [Medline: 26801630]
38. Alexander GL. Issues of trust and ethics in computerized clinical decision support systems. Nurs Adm Q 2006;30(1):21-29. [doi: 10.1097/00006216-200601000-00005] [Medline: 16449881]
39. Amoedo A, Martnez-Costa MDP, Moreno E. An analysis of the communication strategies of Spanish commercial music networks on the web: http://los40.com, http://los40principales.com, http://cadena100.es, http://europafm.es and http://kissfm.es. radio journal: international studies in 2009 Feb 01;6(1):5-20 [FREE Full text] [doi: 10.1386/rajo.6.1.5_4]
40. Yun JM, Shin DW, Hwang S, Cho J, Nam YS, Kim JH, et al. Effect of public disclosure on antibiotic prescription rate for upper respiratory tract infections. JAMA Intern Med 2015 Mar 01;175(3):445-447. [doi: 10.1001/jamainternmed.2014.6569] [Medline: 25506784]

Abbreviations

ADE: adverse drug event
API: application programming interface
ATT: attitude
AU: actual use
CFI: comparative fit index
COREQ: Consolidated Criteria for Reporting Qualitative Research Guidelines
EE: effort expectancy
EHR: electronic health record
EMR: electronic medical record
FC: facilitating conditions
GPI: generic product identifier
HIS: hospital information system
ITU: intention to use
MDSS: medication decision support system
PE: performance expectancy

PEOU: perceived ease of use
PU: perceived usefulness
RMSEA: root mean square error of approximation
SEM: structural equation modeling
SI: social influence
SNUBH: Seoul National University Bundang Hospital
TAM: technology acceptance model
TLI: Tucker-Lewis index
UTAUT: unified theory of acceptance and use of technology

Edited by G Eysenbach; submitted 16.03.20; peer-reviewed by J Aarts, S Sarbadhikari; comments to author 23.04.20; revised version received 02.05.20; accepted 14.05.20; published 22.07.20.

Please cite as:
Jung SY, Hwang H, Lee K, Lee HY, Kim E, Kim M, Cho IY
Barriers and Facilitators to Implementation of Medication Decision Support Systems in Electronic Medical Records: Mixed Methods Approach Based on Structural Equation Modeling and Qualitative Analysis
JMIR Med Inform 2020;8(7):e18758
URL: https://medinform.jmir.org/2020/7/e18758
doi:10.2196/18758
PMID:32706717

Review

Appropriateness of Overridden Alerts in Computerized Physician Order Entry: Systematic Review

Tahmina Nasrin Poly[1,2,3], MSc; Md.Mohaimenul Islam[1,2,3], MSc; Hsuan-Chia Yang[1,2,3], MSc, PhD; Yu-Chuan (Jack) Li[1,2,3,4,5], MD, PhD

[1]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

[2]International Center for Health Information Technology (ICHIT), Taipei Medical University, Taipei, Taiwan

[3]Research Center of Big Data and Meta-Analysis, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

[4]Department of Dermatology, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

[5]TMU Research Center of Cancer Translational Medicine, Taipei Medical University, Taipei, Taiwan

Corresponding Author:
Yu-Chuan (Jack) Li, MD, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
No. 250 Wuxing Street
Taipei, 110
Taiwan
Phone: 886 2 27361661 ext 7600
Email: jaak88@gmail.com

Abstract

Background: The clinical decision support system (CDSS) has become an indispensable tool for reducing medication errors and adverse drug events. However, numerous studies have reported that CDSS alerts are often overridden. The increase in override rates has raised questions about the appropriateness of CDSS application along with concerns about patient safety and quality of care.

Objective: The aim of this study was to conduct a systematic review to examine the override rate, the reasons for the alert override at the time of prescribing, and evaluate the appropriateness of overrides.

Methods: We searched electronic databases, including Google Scholar, PubMed, Embase, Scopus, and Web of Science, without language restrictions between January 1, 2000 and March 31, 2019. Two authors independently extracted data and crosschecked the extraction to avoid errors. The quality of the included studies was examined following Cochrane guidelines.

Results: We included 23 articles in our systematic review. The range of average override alerts was 46.2%-96.2%. An average of 29.4%-100% of the overrides alerts were classified as appropriate, and the rate of appropriateness varied according to the alert type (drug-allergy interaction 63.4%-100%, drug-drug interaction 0%-95%, dose 43.9%-88.8%, geriatric 14.3%-57%, renal 27%-87.5%). The interrater reliability for the assessment of override alerts appropriateness was excellent (kappa=0.79-0.97). The most common reasons given for the override were “will monitor” and “patients have tolerated before.”

Conclusions: The findings of our study show that alert override rates are high, and certain categories of overrides such as drug-drug interaction, renal, and geriatric were classified as inappropriate. Nevertheless, large proportions of drug duplication, drug-allergy, and formulary alerts were appropriate, suggesting that these groups of alerts can be primary targets to revise and update the system for reducing alert fatigue. Future efforts should also focus on optimizing alert types, providing clear information, and explaining the rationale of the alert so that essential alerts are not inappropriately overridden.

(JMIR Med Inform 2020;8(7):e15653) doi:10.2196/15653

KEYWORDS

clinical decision system; computerized physician order entry; alert fatigue; override; patient safety

Introduction

Rationale

A computerized provider order entry (CPOE) system is often integrated with a clinical decision support system (CDSS) to reduce patient harm and error rates [1]. A CDSS has immense potential for fostering patient safety and quality of care by reducing the adverse drug effects (ADEs) rate [1-3]. However, the current CDSS generates too many alerts, which are often overridden (approximately 90% to 95%), sometimes inappropriately. Concern related to inappropriate overrides reached a peak [4,5] with recognition of the potential to increase the risk of harm to patients. Multiple studies have reported that a high frequency of clinically irrelevant alerts (repetitive alerts with minimal clinical value), mediocre functionality (minimal integration among various departments and lack of alerts prioritization), and erroneous assessment by physicians are the main reasons for inappropriate overrides [4,6,7]. However, the growing number of inappropriate overrides often silently puts patients at risk of fatal ADEs [8,9].

To date, significant efforts have been taken to make sound clinical decisions and provide high-quality services. Indeed, lower specificity (high false-positives) and ambiguous alert contents (no clear information provided on why alerts were triggered in the systems) are still associated with excessive overrides and alert fatigue [4,10]. A CDSS with higher sensitivity and lower specificity could also contribute to the substantial number of inappropriate alerts [11,12]. Recent findings suggest that applying hard-stop alerts might be an efficient and helpful tool to reduce inappropriate overrides; however, such a tool must be judiciously implemented to achieve improved usability and receptivity of systems [13]. To increase the alert acceptance rate and reduce overrides, a system should be implemented in such a way that enables prioritizing alerts based on grade and potential harm, analyze the physician response, provide clear recommendations, and explain why the alert is triggered [14].

Goal of Investigation

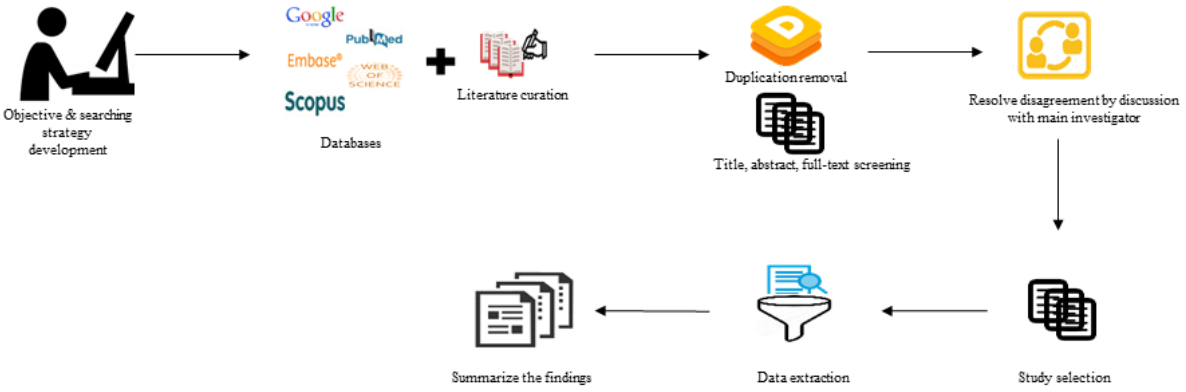
Since the override rate has been increasing, it is necessary to ascertain the types of alerts that are most frequently triggered, calculate the rate at which they are overridden (ie, reject the alerts), and to determine the reasons for overrides and the appropriateness of the reasons. Gaining a better understanding of these issues can provide meaningful insight into how alerts can be delivered in a relevant way (ie, converting a hard alert to a soft alert or turning off clinically irrelevant alerts or those with low clinical value).

Methods

Overview

We conducted a systematic review in accordance with the Meta-analysis of Observational Studies in Epidemiology guidelines [15] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard [16]. The overview of the study process is given in Figure 1.

Figure 1. Overview of the study process.



Electronic Databases Search

We conducted a systematic search in electronic databases, including the PubMed, Embase, Scopus, Google Scholar, and Web of Science databases, between January 1, 2000 and April 30, 2019. The search was performed by two authors (MI and TP) using the keywords “alert fatigue,” “override alerts,” “computerized physician order entry,” “decision support system,” “medication-related CDS,” “CDSS,” and “CPOE.” There was no language and data restriction applied in the initial search. We also scanned the references of review articles and conference proceedings.

Eligibility Criteria

The titles and abstracts of all retrieved studies were screened independently by two expert authors (MI and TP) to find the most relevant articles. They selected potentially eligible full-text articles. The full-text articles were considered as appropriate for inclusion in the systematic review by these same two experts after screening the full text and documenting the reasons for exclusion of inappropriate/ineligible articles. Any disagreement that arose in this screening process was resolved by the principal investigator of the study (YL). Articles were considered for inclusion if they met the following criteria: (1) published in English with desired outcomes reported, (2) evaluated override alerts along with reasons for those overrides, and (3) reported the override rate and the appropriateness of the override reasons.

We excluded studies if they were published in the form of a review, report, short communication, letter to editor, methodology, or editorial.

Data Extraction

For studies that fulfilled the inclusion criteria, two authors (MI and TP) conducted data abstraction using a predefined, standardized protocol. Review Manager software (RevMan-5, Cochrane, UK) was also used to check the accuracy of the included studies. The following information was collected from the included studies: (1) methods, including setting, data analyzed, study design, study period, type of alerts, appropriateness criteria, inclusion and exclusion criteria; (2) results, including number of alerts, percent of override alerts, percentage of different types of alert overrides, percentage of overall override alerts, reasons for those overrides, characteristics of alert types, rate of appropriateness, rate of appropriateness for each override alert subtype, and rate of adverse effects; and (3) discussion, including the main findings, suggestions, intended recommendations, and limitations.

Outcome Parameters

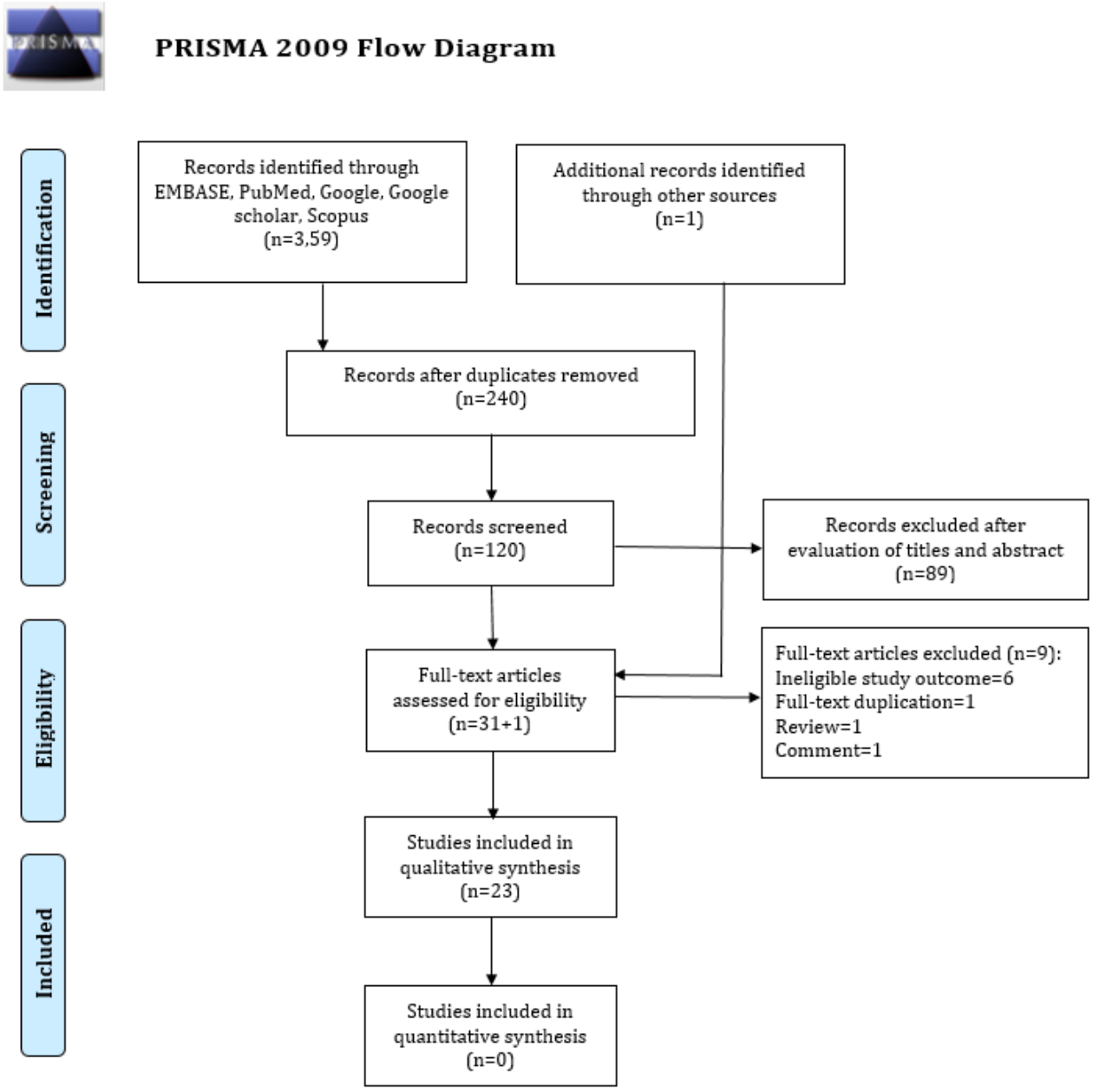
The following three primary outcomes were considered in our analysis: (1) characterize the types of alerts and their override rate; (2) the reasons for an override for different types of alerts assessed for inpatient and outpatient settings; and (3) the rate of the appropriateness of the override reasons.

Results

Literature Selection

Our systematic search identified 360 titles and abstracts of potentially eligible studies for inclusion. Of these, 240 articles were excluded due to duplication and 88 of the remaining 120 articles were excluded based on predefined eligibility criteria during screening of titles and abstracts. The remaining 32 articles were processed for full-text review. Among these, a total of 23 relevant studies met all inclusion criteria [4,5,7,8,11,17-34]. Figure 2 shows all inclusion and exclusion criteria based on the PRISMA guidelines.

Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for study selection.



Study Characteristics

The study characteristics of the included 23 articles are presented in Table 1. In this systematic review, six studies were based on a retrospective observational study design, five studies were cross-sectional, five studies were prospective observational studies, and seven studies only mentioned an observational study design. The settings included the intensive care unit (n=4), primary teaching hospital (n=5), academic medical center (n=5), tertiary care teaching hospital (n=3), university pediatric hospital (n=2), and others (n=4). There were 12 different types of alerts (drug-allergy, drug-drug interaction, drug-class, class-class, drug-dose, drug-duplication, drug-laboratory, drug-disease, drug-pregnancy, geriatric, age-based suggestion, renal, and formulary substitution) discussed in the included studies. The maximum override rate was 96.2%. ADEs were evaluated in 5 of the 23 studies [7,20,21,32,33].

Table 1. Characteristics of included studies.

Reference	Design	Setting	Period	Alert type	Override (%)	ADEs[a] due to inappropriate override
Wong et al [7]	POS[b]	ICU[c]	September 2016-April 2017	Dose-range	93	Increased
Wong et al [17]	POS	In- and outpatients	January 2009-December 2011	DAI[d]	Inpatients, 46; outpatients, 68.8	N/A[e]
Cho et al [18]	ROS[f]	TAH[g]	September 2014-December 2014	DDI[h]	89.4	N/A
Nanji et al [19]	CSS[i]	TCTI[j]	2009-2012	DAI, DDI, DD[k], ABR[l], RR[m], FS[n]	46.2	N/A
Wong et al [20]	POS	ICU	July 2016-April 2017	DAI, DDI, geriatric, renal	88.5	Increased
Rehr et al [8]	POS	ICU	June 2016-November 2016	Dose, DDI, DAI	66.0	N/A
Wong et al [21]	ROS	ICU	2009-2011	DAI, DDI, geriatric, renal	~87.1	Increased
Slight et al [22]	CSS	TCTI	January 2009-December 2011	DAI	81.0	N/A
Topaz et al [23]	RCSS[o]	AMC[p]	2004-2013	DAI	87.6	N/A
Her et al [24]	OS[q]	AMC	January 2012-December 2012	NFM[r]	~61.2	N/A
Topaz et al [25]	OS	AMC	2004-2013	DAI	89.7	N/A
Straichman et al [5]	OS	AMC	November 2013-December 2013	Dose, RDA[s], DT[t], DDI	96.2	N/A
Ahn et al [26]	ROS	ED[u] and GW[v]	September 2009-June 2013	DDI	ED: 94 GW: 57	N/A
Nanji et al [4]	OS	OP[w] and AHBP[x]	Jan 2009-December 2011	DAI, DDI, DD, DCI[y], CCI[z], ABS[aa], RS[bb], FS	52.6	N/A
Cho et al [27]	CSS	OP	January 2009-December 2011	Renal dose	78.2	N/A
Bryant et al [28]	ROS	PTH[cc]	June 10-13, 2013	DDI	~95.1	N/A
Jani et al [34]	ROS	AMC	October 2005-October 2006	DAI, DDI, DT	89.0	N/A
Slight et al [11]	CSOS[dd]	PTH	January 2009-December 2011	DDI	53.4	N/A
Mille et al [29]	POS	UPH[ee]	November 2006-December 2006	DDI	68.7	N/A
Van der Sijs et al [30]	ROS	UPH	2001-2005	DDI	72.0	N/A
Shah et al [31]	OS	PTH	August 2004-January 2005	DD, DDI, DL[ff], DID[gg], DP[hh]	71	N/A
Hsieh et al [32]	OS	PTH	August 2002-October 2002	DAI	80.0	Increased
Weingart et al [33]	OS	PTH	October 2000-December 2000	DDI, DAI	~91.2	Increased

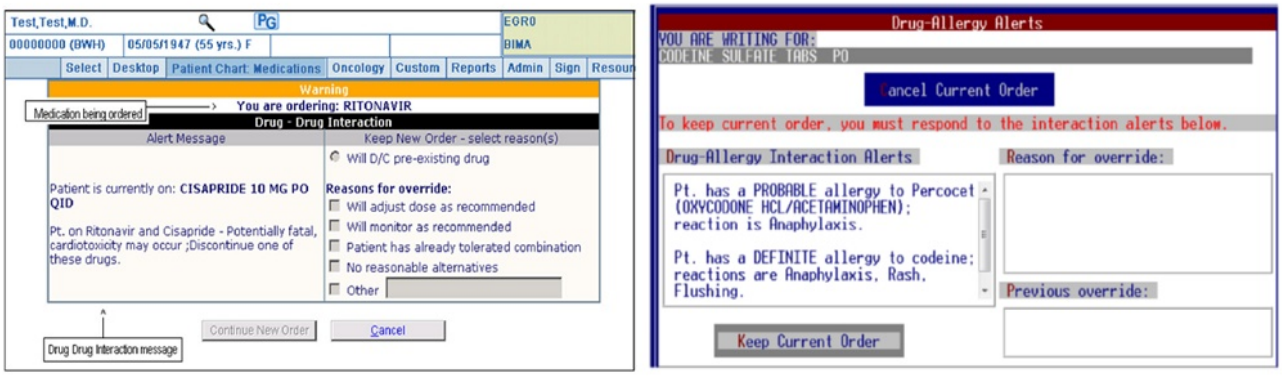
[a]ADE: adverse drug effect.

[b]POS: prospective observational study.
[c]ICU: intensive care unit.
[d]DAI: drug-allergy interaction.
[e]N/A: not applicable.
[f]ROS: retrospective observational study.
[g]TAH: tertiary academic hospital.
[h]DDI: drug-drug interaction.
[i]CSS: cross-sectional study.
[j]TCTI: tertiary-care teaching hospital.
[k]DD: duplicate drug.
[l]ABR: age-based recommendation.
[m]RR: renal recommendation.
[n]FS: formulary substitution.
[o]RCSS: retrospective cross-sectional study.
[p]AMC: academic medical center.
[q]OS: observational study.
[r]NFM: nonformulary medication.
[s]RDA: renal dose adjustment.
[t]DT: duplicate therapy.
[u]ED: emergency department.
[v]GW: general ward.
[w]OP: outpatients.
[x]AHBP: ambulatory hospital-based practice.
[y]DCI: drug-class interaction.
[z]CCI: class-class interaction.
[aa]ABS: age-based suggestion.
[bb]RS: renal suggestion.
[cc]PTH: primary teaching hospital.
[dd]CSOS: cross-sectional observational study.
[ee]UPH: university pediatric hospital.
[ff]DL: drug lab.
[gg]DID: drug-disease.
[hh]DP: drug pregnancy.

Appropriateness Criteria

All of the included studies developed criteria for evaluating the appropriateness of overrides for each alert type for both inpatient and outpatient settings. To validate the appropriateness framework, they used a chart along with previously published articles and clinical experience of the multidisciplinary group (physicians, pharmacists, and nurses). All studies used specific criteria for different types of alerts, which were modified until reaching a final agreement. They considered override alerts as appropriate if the reasons reported by the physicians were acceptable according to their study's framework and also verified based on review of relevant guidelines. For example, if a clinician prescribed a drug and a dose alert was displayed, the appropriate override reasons mentioned were "will monitor as recommended," "will adjust the dose," and "patient has already tolerated" based on previous data, indicating that monitoring is beneficial to patients (Figure 3). Subsequently, the multidisciplinary group carefully checked and verified all of the override reasons in their chart review. They extensively verified by reviewing guidelines, such as checking for dose/renal function/drug monitoring criteria, previously prescribed tolerable medication combinations, and accepted/refused medication. The included studies mentioned that pharmacists, nurses, training health care personal, and clinicians checked and verified the appropriateness of override reasons. Any disagreements among them were resolved with discussion.

Figure 3. Example of a drug-drug interaction alert (left) and drug-allergy interaction alerts (right).



Override Rate and Appropriateness of Overrides

All 23 studies described the alert override rate and the appropriateness of overrides according to alert type (drug-allergy, drug-drug, dose, drug-class, class-class, drug duplication, drug-laboratory, drug-disease, drug-pregnancy, geriatric, renal-dose, age-based suggestion, renal, formulary substitution). The average override alerts ranged from 46.2% to 96.2% (Table 1). However, the range of override rates varied according to alert type (drug-allergy 46%-95%, drug-drug interaction 56.3%-95.6%, dose 82%-96.8%, geriatric 2.1%-87.1%, and renal 74.4%-97.1%). Moreover, the overall appropriateness rate ranged from 29.4% to 100%, which also varied according to alert types (drug-allergy 63.4%-100%, drug-drug interaction 0%-95%, dose 43.9%-88.8%, geriatric 14.3%-57%, renal 27%-87.5%). However, interrater reliability for the assessment of override alerts appropriateness was excellent (kappa=0.79-0.97). Table 2 summarizes the rates of override alerts and the appropriateness of override alerts by type.

Table 2. Rate of override alerts and appropriateness of override alerts by type.

Reference	Type of alerts	Overridden (%)	Appropriateness (%)	Evaluation criteria	Evaluation rater	Interrater agreement, kappa (95% CI)
Wong et al [7]	Dose	93.0	88.8	Based on previously published data including guidelines	Clinical pharmacist and research assistant	0.87 (0.85-0.90)
Wong et al [17]	DAI[a]	Inpatient: 46.0; outpatient: 68.8	Inpatient: 83.9; outpatient: 100	NR[b]	NR	NR
Cho et al [18]	DDI[c]	71.7	~75.3	Based on previously published data including guidelines	Physicians	0.92
Nanji et al [19]	DAI, DDI, and DD[d]	DAI: 81.9, DDI: 68.2, DD: 51.9	DAI: 96.5, DDI: 62.0, DD: 98.0	Based on previously published data including guidelines	Physician and pharmacist	0.96 (0.95-0.97)
Wong et al [20]	DAI, DDI, dose, geriatric, renal	DAI: 83.6, DDI: 91.9, dose: 96.8, geriatric: 2.30, renal: 97.1	DAI: 83.4, DDI: 82.0, dose: 43.9, geriatric: 14.3, renal: 87.5	Based on previously published data including guidelines	Clinical pharmacist	0.89 (0.85-0.93)
Rehr et al [8]	DAI, DDI, dose	DAI: ~80, DDI: ~87, dose: ~82	DAI: 83.0, DDI: 0.00, dose: 85.0	NR	NR	NR
Wong et al [21]	DAI, DDI, geriatric, renal	DAI: 46.3, DDI: 56.3, geriatric: 87.1, renal: 74.4	DAI: 94.0, DDI: 84.0, geriatric: 57.0, renal: 27.0	Based on previously published data including guidelines	Clinical pharmacist	0.79 (0.73-0.86)
Slight et al [22]	DAI	Inpatient: 83.0; outpatient: 81.0	Inpatient: 96.5; Outpatient: 94.0	Based on previously published data including guidelines	Pharmacist	0.86
Topaz et al [23]	DAI	~87.6	NR	NR	NR	NR
Her et al [24]	FA[e]	~61.2	82.8	Based on previously published data including guidelines	Pharmacist	0.97 (0.92-1.00)
Topaz et al [25]	DAI	89.7	NR	NR	NR	NR
Straichman et al [5]	Dose, RDA[f], DT[g], DDI, MDDI[h]	Dose, 92.1; RDA: 92.3, DT: 96.0, DDI: 95.6, MDDI: 96	Overall: 84.5	Based on previously published data including guidelines	Pharmacist	NR
Ahn et al [26]	DDI (ED[i]), DDI (GW[j])	DDI (ED): 94.0, DDI (GW): 57.0	DDI (ED): 59.6, DDI (GW): 40.4	NR	NR	NR
Nanji et al [4]	DAI, DDI, DD[f], DCI[k], CCI[l], ABS[m], renal, FA	DAI: 77.4, DDI: 60.2, DD: 28.6, DCI: 24.4, CCI: 69.7, ABS: 79.0, renal: 78.0, FA: 85.0	DAI: 92.0, DDI: 12.0, DD: 82.0, DCI: 88.0, CCI: 69.0, ABS: 39.0, renal: 12.0, FA: 57.0	Based on previously published data including guidelines	Physician, pharmacist, and nurse	0.89
Cho et al [27]	Renal	78.2	29.4	Based on previously published data including guidelines	Physician	0.93
Jani et al [34]	DAI, DDI, DT, ED	DAI: 63.4, DDI: 73.0, DT: 95.0, ED: 90.6	NR	NR	NR	NR
Bryant et al [28]	DDI and DAI	DDI: 95.0 and DAI: 91.0	NR	NR	NR	NR
Slight et al [11]	DDI	60.0	68.2	Based on previously published data including guidelines	Pharmacist	0.84
Mille et al [29]	DDI	68.7	NR	NR	NR	NR

Reference	Type of alerts	Overridden (%)	Appropriateness (%)	Evaluation criteria	Evaluation rater	Interrater agreement, kappa (95% CI)
Van der Sijs et al [30]	DDI	72.0	NR	NR	NR	NR
Shah et al [31]	DCI, DDI, DLI[n], DRDI[o], and DPI[p]	DCI: 23.0, DDI: 58.0, DLI: 60.0, DRDI: 47.0, DPI: 90.0	NR	NR	NR	NR
Hsieh et al [32]	DAI	80.0	NR	NR	NR	NR
Weingart et al [33]	DAI and DDI	DAI: 91.2; DDI: 94.6	63.5	Based on previously published data including guidelines	Board-certified internee	0.86

[a]DAI: drug-allergy interaction.
[b]NR: not reported.
[c]DDI: drug-drug interaction.
[d]DD: drug duplicate.
[e]FA: formulary alert.
[f]RDA: renal dose adjustment.
[g]DT: duplicate therapy.
[h]MDDI: major drug-drug interaction.
[i]ED: emergency department.
[j]GW: general ward.
[k]DCI: drug-class interaction.
[l]CCI: class-class interaction.
[m]ABS: age-based suggestion.
[n]DLI: drug-lab interaction.
[o]DRDI: drug-disease interaction.
[p]DPI: drug-pregnancy interaction.

Reasons for Overrides

All 23 included studies evaluated the reasons for overriding the alerts in the CDSS. The most common reasons for overriding drug-allergy alerts were “will monitor,” “patients tolerated before,” “patient took previously without an allergic reaction,” “low risk across sensitivity,” “no reasonable alternatives,” and other (with or without a free-text reason provided). The most common reasons for overriding drug-drug interaction alerts were “will monitor as recommended,” “will adjust the dose as recommended,” “patients have already tolerated this combination,” “clinically irrelevant,” and “benefit assessed to be greater than the risk.” Moreover, the common reasons for overriding dose alerts were “will adjust the dose as recommended,” “benefit outweighs risk,” and “patients tolerated before” (Table 3).

Table 3. Override reason by alert type.

Alert type	Override reason
Drug-allergy	Will monitor [4,8,19,20,25,32,34].
	Patient tolerated before [8,20,21,25,32-34].
	Patient took previously without allergic reaction [4,5,17,19-22,32,34].
	Provider approved [28].
	Low risk across sensitivity [4,19,21,22].
	No reasonable alternatives [4,22,25,33].
	Limited course of treatment [33].
	Physician aware [17,19,21,22,32,34].
	Alerted interaction not clinically significant [28].
	Allergy information inaccurate in patient's records [33].
	Patient does not have this allergy, will D/C the pre-existing allergy [17,25,32].
	Desensitization [17].
	Administer per desensitization protocol [17].
	Other (allows user to enter free text) [8,22,25,30,33].
	Other (with no free-text reason provided) [4,22,33].
	Unknown [21,23,25,30].
Drug-drug interaction	Will monitor as recommended [4,5,8,11,19-21,31,34].
	Will adjust dose as recommended [4,8,11,19,21,31].
	Patient has already tolerated this combination [5,8,11,19,21,31,33,34].
	No reasonable alternative [4,11,31,33].
	Clinically irrelevant alert [5,26,28,30,33].
	Medication list out of date [33].
	Limited course of treatment [4,33].
	Benefit assessed to be greater than the risk [5,26,29,33].
	The drug combination will be given only for a short period and is therefore safe [5].
	The computerized system did not interpret my prescription correctly [5].
	The drug-drug interaction is unlikely to occur because of the route of administration [5].
	Combinations of the coded reasons listed above [11,30].
	Other (allows user to enter free text) [4,11,26,30,33].
	Other (with no free-text reason provided) [4,11,19,33].
Drug-class	Will monitor as recommended [4].
	Will adjust the dose as recommended [4].
	Patient has already tolerated this combination [4].
	No reasonable alternatives [4].
	Others [4].
Class-class	Will monitor as recommended [4].
	Will adjust the dose as recommended [4].
	Patient has already tolerated this combination [4].
	No reasonable alternatives [4].
	Others [4].
Drug-dose	Will adjust dose as recommended [5,7,8,20].
	Benefit outweighs risk [5,7].
	Patients tolerated before [5,7].
	Inaccurate warning [7,8].
	The drug combination or the drug at the given dose before without any adverse effects [5].
	The drug dose alert is based on patient weight which is unavailable in the electronic patient record [5].

Alert type	Override reason
Drug-duplication	Combination therapy indicated [19].
	One-time dose [19].
	Not duplicate therapy [19].
	Patient requires different strengths of the same drug [4,5].
	Transitioning from one drug to the other [4,31].
	Patient on long-term therapy with a combination [4,31].
	Advice from a consultant [4].
	New evidence supports duplicate therapy of this type [4].
	Others [4].
Drug-lab	Will monitor/manage as recommended [31].
	More recent lab results available that warrant use [31].
Drug-disease	Patient has tolerated this drug in the past [31].
	New evidence supports the therapy of this type [31].
Drug-pregnancy	Patient is not pregnant [31].
	Patient is not of child-bearing potential [31].
	Advice from a consultant [31].
	No reasonable alternative [31].
	Patient has tolerated in the past [31].
	Medication is for short-term/as-needed use only [31].
Geriatric	Patient tolerated before [21].
	Will monitor later [20].
Age-based suggestion	Patient has tolerated this drug in the past [4].
	Advice from a consultant [4].
	New evidence supports the therapy of this type [4].
	Others [4].
Renal suggestion	Will monitor as recommended [5,20,21].
	Patient has tolerated this drug in the past [4,5,21,27].
	New evidence supports the therapy of this type [4,27].
	Advice from a consultant [4,27].
	The computerized system did not interpret my prescription correctly [5].
	Others [4,27].
Formulary substitution	Intolerance/failure of suggested substitution [4,24].
	Patient preference [4].
	Patients currently taking prescribed medication [4].
	Insurance does not allow the above suggestion [4].
	Written originally by another physician [4].
	Pharmacological [24].
	Specialist recommendation [24].
	Disease or condition [24].
	Blank [24].
	Others [4].

ADEs

Five studies compared ADEs based on an appropriate and inappropriate override. Wong et al [17] demonstrated a significantly increased rate of ADEs in inappropriate override dose alerts compared with appropriate override dose alerts. The rate of ADEs per 100 override dose alerts was 1.3 and 5.0 for appropriate and inappropriate override dose alerts, respectively. Wong et al [20] also evaluated the potential and definite ADEs in 5 different types of alerts reported, demonstrating that the average potential and definite ADEs were higher in alerts that were considered to be inappropriately overridden than appropriate override alerts (16.5 vs 2.74 per 100 overridden alerts, $P<.001$). However, the rate of ADEs was always higher for inappropriate override alerts (drug-allergy interaction: 11.5 vs 0.6; drug-drug interaction: 11.4 vs 2.0; dose: 17.6 vs 11.1; geriatric: 11.1 vs 0; and renal: 30.8 vs 0). The logistic regression model showed that inappropriate override alerts were significantly associated with an increased risk of ADEs (odds

ratio 6.14, 95% CI 4.63-7.71, *P*<.001) and an increased intensive care unit length of stay (2.25 days, 95% CI 0.52-3.98, *P*=.01). Moreover, 3 studies also reported that inappropriate override was associated with an increased risk of ADEs [21,32,33].

Discussion

Main Findings

This is the first systematic review that evaluated the current scenario of a CDSS by measuring the rate with which alerts are overridden, described the reasons for override alerts at the time of prescribing, and the appropriateness of overrides. A significant proportion of alerts in the CDSS were overridden (96.2%), and the override rate varied dramatically according to alert types. The rate of appropriate overrides was high (nearly 100%) and they also varied significantly according to alert types. For example, renal, geriatric, and drug-drug interaction alert overrides had low appropriateness rates, whereas drug-allergy, drug-duplication, drug-formulary, and drug class alert overrides had higher appropriateness rates. Inappropriate overrides were associated with an increased risk of ADEs when compared with appropriately overridden alerts. However, the reasons provided for overriding alerts varied extensively depending on alert types. Refinement of these alert types has immense potential to improve the acceptance rate and patient safety. Furthermore, the clinical team should evaluate the appropriateness of overrides based on the given clinical context to optimize alert types and frequencies and ultimately improve their clinical relevancy while reducing alert fatigue.

Clinical Implications

A CPOE integrated with a CDSS is designed to improve patient safety and reduce preventable errors by generating pop-up alerts at the point of order entry. A frequent complaint about CPOEs is firing up too many alerts, which are frequently not clinically relevant or have very low clinical value [35]. An excessive number of alerts in the CPOE desensitizes physicians (hampers the mental state, consumes too much time), leading them to override both appropriate and clinically irrelevant alerts [36]. A system with low sensitivity (ie, more false-negatives) and low specificity (ie, more false-positives), ambiguous information content, and an overwhelming number of alerts (both relevant and irrelevant alerts) induce alert fatigue [35]. Inappropriate override always leads to potential ADEs and increased morbidity [37]. A study evaluating drug-drug interaction alert overrides and how override alerts lead to preventable ADEs reported 22 serious ADEs over the 3-month study period [32].

In our study, a high number of alerts were overridden, especially for dose, drug-drug interaction, and drug-allergy interaction alerts. The findings of our study also suggest that a higher number of these alerts can lead to alert fatigue. There are two ways to combat alert fatigue. First, the system should set a higher threshold for triggering alerts. Second, the most frequent alerts should be categorized and the system updated regularly (overrides tend to increase over time). We also evaluated the appropriateness of alert overrides, demonstrating that the rate of the appropriateness of overrides varied according to the different types of alerts. Evaluating the appropriateness of override alerts is difficult but the range of interrater reliability for assessment was high. Among the dose recommendation alerts that were overridden, only 43.9%-85% were found to be appropriately overridden. The range of appropriately overridden renal and drug-drug interaction alerts was 12%-87.5% and 0%-84%, respectively. Among the overridden drug-allergy interaction recommendation alerts, approximately 83.5%-100% were appropriately overridden. Moreover, the vast majority of drug-duplicate (82%-99%), drug-class (88%), and formulary (82.8%) override alerts were appropriate, indicating that these groups can be the primary targets for rectification to stop alert fatigue by reducing or converting (hard-stop alert to soft/passive alerts) the number of alerts. However, the higher rates of inappropriate overrides of the renal, drug-drug interaction, and geriatric alert types indicate the need for further intervention.

Our findings also provide a variety of reasons for overriding alerts. The majority of physicians provided the reasons for overrides as “will monitor as recommended,” “patients have tolerated it before,” “will adjust the dose,” and “maximum time,” leaving the free-text box blank. In some cases, physicians do not write any reason for the overrides; however, it is important to clearly outline the override reasons to best invest in the patient’s condition and care. Indeed, these findings raise concern about patient safety and quality of care. For example, failure to monitor several drug levels such as digoxin after initiation of verapamil (drug-drug interaction) can cause serious harm for the patient [38]. Other common reasons physicians gave for the override were “no reasonable alternative,” “physician aware,” “patients have already tolerated this combination,” and other (without free-text reason provided). However, there was no confirmation that physicians were actually aware of the potential harm and had monitored the patient’s condition before overriding. Moreover, in the case of drug-allergy, approximately two-thirds of alerts that showed a reaction of “anaphylaxis” were overridden by physicians and with the reasons provided including “patients have taken previously without an allergic reaction” and “low-risk cross-sensitivity.” However, a patient with a true allergy can experience severe anaphylaxis. For example, a reaction between vancomycin and red man syndrome was found to be inappropriately overridden with the reason given that the “patient has taken previously without allergic reaction/patient has tolerated previously,” but severe ADE was observed (development of a patchy macular rash) [21]. Therefore, it is essential to know the patient’s history of anaphylaxis to reduce serious recurrence (approximately 35% of patients experience recurrence) [39]. Topaz et al [23] reported that only about one-tenth of the alerts showed potential life-threatening effects that were a definite match between the allergy and prescribed drugs, although others were due to either the “cross-sensitivity or allergy group.” Several studies confirmed that the hospitalization of patients with anaphylaxis has been increasing in both the United States [40] and the United Kingdom [41]. It is therefore important to evaluate these types of override alerts and the reasons given by the physician for the override. Moreover, future studies are needed to develop an effective knowledge management system that can provide more accurate and relevant drug-allergy interaction alerts for improving patient safety.

Recommendations to Improve the CDSS

The findings of our study provide a clear picture of the overall situation of current CDSSs by summarizing the existing literature. These findings can help policymakers and researchers to improve existing CDSSs by conducting an in-depth analysis of existing CDSS features. Having provided a collection of evidence-based information and removing unimportant alerts, a novel system also requires rigorous evaluation to determine the optimum rate of sensitivity and specificity for reducing patient harm. No system can achieve 100% sensitivity and specificity in a real-world setting. However, a logical and effective symmetry between sensitivity and specificity can make the system more flexible and safer. The sensitivity and specificity should be increased without sacrificing the other through the combination of patient factors and using futuristic algorithms. Osheroff et al [42] demonstrated that "five rights" (right information, to the right person, in the right CDS intervention format, through the right channel, and at the right time in the workflow) should be taken into consideration when alerts will be popped up in the system. Several recommendations are provided below to design a sophisticated CDSS by reducing alert fatigue.

First, increase the positive predictive value for dose recommendation alerts by incorporating patient-specific factors (eg age, other medication orders, renal impairment history) [7].

Second, optimize alert types and frequencies to increase their clinical relevance so that important alerts are not inappropriately overridden [43].

Third, override alerts can be revised if they are not clinically important, and the system will be updated for reducing alert fatigue.

Fourth, turn off alerts that are not clinically important/inaccurate or of only minor importance [8,44].

Fifth, it is essential to categorize the most frequent interruptive alerts; for serious alerts such as drug-drug interactions and renal, the dose should be displayed as interruptive, whereas minor/low-risk alerts can be presented in a noninterruptive manner [13].

Sixth, all types of alerts should contain clear and concise information [45,46] and provide exact information on why the alert is important for the situation [47].

Seventh, identify a list of medications that patients previously showed no allergic reaction to or tolerated in the past so that physicians are not inundated with highly irrelevant alerts. Alerts to previously tolerated medications might be presented in a noninterruptive fashion [48-50].

Eighth, systems should pay more attention to the storage of override reasons data (eg, dose-range, allergy) [50,51], and encourage providers to provide accurate override reasons [52,53].

Ninth, identify the malfunctions and pattern of malfunctions in the CDSS [54].

Tenth, it is essential to remove the repetitive and duplicate nature of alerts in the CDSS [21,55].

Eleventh, it is important to understand the system behavior and patterns of physicians in accepting and rejecting the alerts [18].

Twelfth, the system can trigger an alert based on the specialty of physicians (eg, do not provide too many renal alerts for kidney specialists and those with many years of experience in this field) [30,56].

Thirteenth, a drug-drug interaction alert can be presented in an "alert tiering" based on the level of severity. For example, level-1, level-2, and level-3 will be considered as life-threatening, less serious, and least serious, respectively. For level-1 and level-2, hard-stop alerts will be applied, whereas a passive alert (no need for physicians action) can be applied [57].

Fourteenth, it is important to use hard-stop alerts for drug-drug interactions, renal, and geriatric alerts that might harm patients and to use soft-stop alerts for a formulary, drug-allergy, and drug-class alerts that have a lower risk for patient harm [57].

Fifteenth, review alerts periodically and improve according to clinical importance [58-60].

Sixteenth, always encourage physicians to provide override reasons. Learn from the override reasons and place maximum effort to improve the system [61].

Seventeenth, when designing the system, form a multidisciplinary committee consisting of physicians, pharmacists, information technology specialists, and quality administrators [62-64].

Finally, do not establish a silo alert system (always integrate multidepartment data) [62].

Strengths and Limitations

There are several strengths of our study that should be mentioned. First, this is the very first systematic review that summarizes the overall override rate, reasons for overriding the alerts, and the appropriateness of the reasons. Second, we have also provided an override rate and the proportion of appropriateness according to various types of alerts. These data can help policymakers in determining the area that they should place more focus to reduce alert fatigue. Finally, we have provided recommendations to optimize alert types and to improve the clinical relevance of alerts while suppressing alert fatigue for the CDSS that is often injudiciously overridden.

Our study also has several limitations. First, we could not determine the bias of the included studies because of the heterogeneous nature of the studies. Second, we could not provide the percentage of ADEs when alerts were inappropriately overridden owing to data scarcity. Finally, some studies used a random sampling method of alert overrides reviewed for appropriateness that was very trivial compared with the entire alerts fired up in the CDSS, and the accuracy of such reviews was completely reliant on information contained in the patients' charts. However, this may vary from study to study.

Conclusion

The findings of our study show that a high proportion of alerts are overridden and the rate of appropriateness varies widely

according to alert type. Although the CDSS is an extremely effective tool for reducing patient harm and improving quality of care, it could also diminish patient safety if information technology vendors and health care professionals do not appropriately design the clinical interface. Future research should be focused on how to obtain meaningful information for analyzing these override reasons and how to integrate patient-specific factors to reduce alert fatigue, resulting in a more efficient, safe, and effective system.

Acknowledgments

We would like to thank our colleague who is a native English speaker for editing our manuscript. This research is funded in part by the Ministry of Education (MOE) under grants MOE 108-6604-001-400 and DP2-109-21121-01-A-01, and by the Ministry of Science and Technology (MOST) under grants MOST 108-2823-8-038-002 and 109-2222-E-038-002-MY2.

Conflicts of Interest

None declared.

References

1. Wright A, McEvoy DS, Aaron S, McCoy AB, Amato MG, Kim H, et al. Structured override reasons for drug-drug interaction alerts in electronic health records. J Am Med Inform Assoc 2019 Oct 01;26(10):934-942 [FREE Full text] [doi: 10.1093/jamia/ocz033] [Medline: 31329891]
2. Berner E. Agency for Healthcare Research and Quality. 2009 Jun. Clinical decision support systemstate of the art (AHRQ Publication No 09-0069-EF) URL: https://digital.ahrq.gov/sites/default/files/docs/page/09-0069-EF_1.pdf [accessed 2019-05-08]
3. Burgos F, Melia U, Vallverdú M, Velickovski F, Lluch-Ariet M, Caminal P, et al. Clinical decision support system to enhance quality control of spirometry using information and communication technologies. JMIR Med Inform 2014;2(2):e29 [FREE Full text] [doi: 10.2196/medinform.3179] [Medline: 25600957]
4. Nanji KC, Slight SP, Seger DL, Cho I, Fiskio JM, Redden LM, et al. Overrides of medication-related clinical decision support alerts in outpatients. J Am Med Inform Assoc 2014;21(3):487-491 [FREE Full text] [doi: 10.1136/amiajnl-2013-001813] [Medline: 24166725]
5. Zenziper Straichman Y, Kurnik D, Matok I, Halkin H, Markovits N, Ziv A, et al. Prescriber response to computerized drug alerts for electronic prescriptions among hospitalized patients. Int J Med Inform 2017 Nov;107:70-75. [doi: 10.1016/j.ijmedinf.2017.08.008] [Medline: 29029694]
6. Lin C, Payne TH, Nichol WP, Hoey PJ, Anderson CL, Gennari JH. Evaluating clinical decision support systems: monitoring CPOE order check override rates in the Department of Veterans Affairs' Computerized Patient Record System. J Am Med Inform Assoc 2008;15(5):620-626 [FREE Full text] [doi: 10.1197/jamia.M2453] [Medline: 18579840]
7. Wong A, Rehr C, Seger DL, Amato MG, Beeler PE, Slight SP, et al. Evaluation of Harm Associated with High Dose-Range Clinical Decision Support Overrides in the Intensive Care Unit. Drug Saf 2019 Apr;42(4):573-579. [doi: 10.1007/s40264-018-0756-x] [Medline: 30506472]
8. Rehr CA, Wong A, Seger DL, Bates DW. Determining Inappropriate Medication Alerts from "Inaccurate Warning" Overrides in the Intensive Care Unit. Appl Clin Inform 2018 Apr;9(2):268-274 [FREE Full text] [doi: 10.1055/s-0038-1642608] [Medline: 29695013]
9. Peterson JF, Bates DW. Preventable medication errors: identifying and eliminating serious drug interactions. J Am Pharm Assoc (Wash) 2001;41(2):159-160. [doi: 10.1016/s1086-5802(16)31243-8] [Medline: 11297326]
10. Seidling HM, Phansalkar S, Seger DL, Paterno MD, Shaykevich S, Haefeli WE, et al. Factors influencing alert acceptance: a novel approach for predicting the success of clinical decision support. J Am Med Inform Assoc 2011;18(4):479-484 [FREE Full text] [doi: 10.1136/amiajnl-2010-000039] [Medline: 21571746]
11. Slight SP, Seger DL, Nanji KC, Cho I, Maniam N, Dykes PC, et al. Are we heeding the warning signs? Examining providers' overrides of computerized drug-drug interaction alerts in primary care. PLoS One 2013;8(12):e85071 [FREE Full text] [doi: 10.1371/journal.pone.0085071] [Medline: 24386447]
12. Coleman JJ, van der Sijs H, Haefeli WE, Slight SP, McDowell SE, Seidling HM, et al. On the alert: future priorities for alerts in clinical decision support for computerized physician order entry identified from a European workshop. BMC Med Inform Decis Mak 2013 Oct 01;13:111 [FREE Full text] [doi: 10.1186/1472-6947-13-111] [Medline: 24083548]
13. Powers EM, Shiffman RN, Melnick ER, Hickner A, Sharifi M. Efficacy and unintended consequences of hard-stop alerts in electronic health record systems: a systematic review. J Am Med Inform Assoc 2018 Nov 01;25(11):1556-1566 [FREE Full text] [doi: 10.1093/jamia/ocy112] [Medline: 30239810]
14. McCoy AB, Waitman LR, Lewis JB, Wright JA, Choma DP, Miller RA, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc 2012;19(3):346-352 [FREE Full text] [doi: 10.1136/amiajnl-2011-000185] [Medline: 21849334]

15. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000 Apr 19;283(15):2008-2012. [doi: 10.1001/jama.283.15.2008] [Medline: 10789670]
16. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-269. [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
17. Wong A, Seger DL, Slight SP, Amato MG, Beeler PE, Fiskio JM, et al. Evaluation of 'Definite' Anaphylaxis Drug Allergy Alert Overrides in Inpatient and Outpatient Settings. Drug Saf 2018 Mar;41(3):297-302. [doi: 10.1007/s40264-017-0615-1] [Medline: 29124665]
18. Cho I, Lee Y, Lee J, Bates DW. Wide variation and patterns of physicians' responses to drug-drug interaction alerts. Int J Qual Health Care 2019 Mar 01;31(2):89-95. [doi: 10.1093/intqhc/mzy102] [Medline: 29741633]
19. Nanji KC, Seger DL, Slight SP, Amato MG, Beeler PE, Her QL, et al. Medication-related clinical decision support alert overrides in inpatients. J Am Med Inform Assoc 2018 May 01;25(5):476-481. [doi: 10.1093/jamia/ocx115] [Medline: 29092059]
20. Wong A, Amato MG, Seger DL, Rehr C, Wright A, Slight SP, et al. Prospective evaluation of medication-related clinical decision support over-rides in the intensive care unit. BMJ Qual Saf 2018 Sep;27(9):718-724. [doi: 10.1136/bmjqs-2017-007531] [Medline: 29440481]
21. Wong A, Amato MG, Seger DL, Slight SP, Beeler PE, Dykes PC, et al. Evaluation of medication-related clinical decision support alert overrides in the intensive care unit. J Crit Care 2017 Jun;39:156-161. [doi: 10.1016/j.jcrc.2017.02.027] [Medline: 28259059]
22. Slight SP, Beeler PE, Seger DL, Amato MG, Her QL, Swerdloff M, et al. A cross-sectional observational study of high override rates of drug allergy alerts in inpatient and outpatient settings, and opportunities for improvement. BMJ Qual Saf 2017 Mar;26(3):217-225 [FREE Full text] [doi: 10.1136/bmjqs-2015-004851] [Medline: 26993641]
23. Topaz M, Seger DL, Slight SP, Goss F, Lai K, Wickner PG, et al. Rising drug allergy alert overrides in electronic health records: an observational retrospective study of a decade of experience. J Am Med Inform Assoc 2016 May;23(3):601-608. [doi: 10.1093/jamia/ocv143] [Medline: 26578227]
24. Her QL, Amato MG, Seger DL, Beeler PE, Slight SP, Dalleur O, et al. The frequency of inappropriate nonformulary medication alert overrides in the inpatient setting. J Am Med Inform Assoc 2016 Sep;23(5):924-933. [doi: 10.1093/jamia/ocv181] [Medline: 27002076]
25. Topaz M, Seger DL, Lai K, Wickner PG, Goss F, Dhopeshwarkar N, et al. High Override Rate for Opioid Drug-allergy Interaction Alerts: Current Trends and Recommendations for Future. Stud Health Technol Inform 2015;216:242-246 [FREE Full text] [Medline: 26262047]
26. Ahn EK, Cho S, Shin D, Jang C, Park RW. Differences of Reasons for Alert Overrides on Contraindicated Co-prescriptions by Admitting Department. Healthc Inform Res 2014 Oct;20(4):280-287 [FREE Full text] [doi: 10.4258/hir.2014.20.4.280] [Medline: 25405064]
27. Cho I, Slight SP, Nanji KC, Seger DL, Maniam N, Dykes PC, et al. Understanding physicians' behavior toward alerts about nephrotoxic medications in outpatients: a cross-sectional analysis. BMC Nephrol 2014 Dec 15;15:200 [FREE Full text] [doi: 10.1186/1471-2369-15-200] [Medline: 25511564]
28. Bryant AD, Fletcher GS, Payne TH. Drug interaction alert override rates in the Meaningful Use era: no evidence of progress. Appl Clin Inform 2014;5(3):802-813 [FREE Full text] [doi: 10.4338/ACI-2013-12-RA-0103] [Medline: 25298818]
29. Mille F, Schwartz C, Brion F, Fontan J, Bourdon O, Degoulet P, et al. Analysis of overridden alerts in a drug-drug interaction detection system. Int J Qual Health Care 2008 Dec;20(6):400-405. [doi: 10.1093/intqhc/mzn038] [Medline: 18784269]
30. Moltu C, Stefansen J, Svisdahl M, Veseth M. Negotiating the coresearcher mandate - service users' experiences of doing collaborative research on mental health. Disabil Rehabil 2012;34(19):1608-1616. [doi: 10.3109/09638288.2012.656792] [Medline: 22489612]
31. Shah NR, Seger AC, Seger DL, Fiskio JM, Kuperman GJ, Blumenfeld B, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. J Am Med Inform Assoc 2006;13(1):5-11 [FREE Full text] [doi: 10.1197/jamia.M1868] [Medline: 16221941]
32. Hsieh TC, Kuperman GJ, Jaggi T, Hojnowski-Diaz P, Fiskio J, Williams DH, et al. Characteristics and consequences of drug allergy alert overrides in a computerized physician order entry system. J Am Med Inform Assoc 2004;11(6):482-491 [FREE Full text] [doi: 10.1197/jamia.M1556] [Medline: 15298998]
33. Weingart SN, Toth M, Sands DZ, Aronson MD, Davis RB, Phillips RS. Physicians' decisions to override computerized drug alerts in primary care. Arch Intern Med 2003 Nov 24;163(21):2625-2631. [doi: 10.1001/archinte.163.21.2625] [Medline: 14638563]
34. Jani YH, Barber N, Wong ICK. Characteristics of clinical decision support alert overrides in an electronic prescribing system at a tertiary care paediatric hospital. Int J Pharm Pract 2011 Oct;19(5):363-366. [doi: 10.1111/j.2042-7174.2011.00132.x] [Medline: 21899617]
35. Khajouei R, Jaspers MWM. The impact of CPOE medication systems' design aspects on usability, workflow and medication orders: a systematic review. Methods Inf Med 2010;49(1):3-19. [doi: 10.3414/ME0630] [Medline: 19582333]

36. Cash JJ. Alert fatigue. Am J Health Syst Pharm 2009 Dec 01;66(23):2098-2101. [doi: 10.2146/ajhp090181] [Medline: 19923309]
37. Ohta Y, Sakuma M, Koike K, Bates DW, Morimoto T. Influence of adverse drug events on morbidity and mortality in intensive care units: the JADE study. Int J Qual Health Care 2014 Dec;26(6):573-578. [doi: 10.1093/intqhc/mzu081] [Medline: 25192926]
38. Cañas F, Tanasijevic MJ, Ma'luf N, Bates DW. Evaluating the appropriateness of digoxin level monitoring. Arch Intern Med 1999 Feb 22;159(4):363-368. [doi: 10.1001/archinte.159.4.363] [Medline: 10030309]
39. Mullins RJ. Anaphylaxis: risk factors for recurrence. Clin Exp Allergy 2003 Aug;33(8):1033-1040. [doi: 10.1046/j.1365-2222.2003.01671.x] [Medline: 12911775]
40. Ma L, Danoff TM, Borish L. Case fatality and population mortality associated with anaphylaxis in the United States. J Allergy Clin Immunol 2014 Apr;133(4):1075-1083 [FREE Full text] [doi: 10.1016/j.jaci.2013.10.029] [Medline: 24332862]
41. Turner PJ, Gowland MH, Sharma V, Ierodiakonou D, Harper N, Garcez T, et al. Increase in anaphylaxis-related hospitalizations but no increase in fatalities: an analysis of United Kingdom national anaphylaxis data, 1992-2012. J Allergy Clin Immunol 2015 Apr;135(4):956-963 [FREE Full text] [doi: 10.1016/j.jaci.2014.10.021] [Medline: 25468198]
42. Osheroff J. Improving Medication Use and Outcomes with Clinical Decision Support: A Step by Step Guide. Chicago: HIMSS; 2009.
43. Coleman JJ, Hodson J, Ferner RE. Deriving dose limits for warnings in electronic prescribing systems: statistical analysis of prescription data at University Hospital Birmingham, UK. Drug Saf 2012 Apr 01;35(4):291-298. [doi: 10.2165/11594810-000000000-00000] [Medline: 22263779]
44. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. Health Aff (Millwood) 2011 Dec;30(12):2310-2317. [doi: 10.1377/hlthaff.2010.1111] [Medline: 22147858]
45. Kawamoto K, Lobach DF. Clinical decision support provided within physician order entry systems: a systematic review of features effective for changing clinician behavior. AMIA Annu Symp Proc 2003:361-365 [FREE Full text] [Medline: 14728195]
46. Feldstein A, Simon SR, Schneider J, Krall M, Laferriere D, Smith DH, et al. How to Design Computerized Alerts to Ensure Safe Prescribing Practices. Joint Commis J Qual Safet 2004 Nov;30(11):602-613. [doi: 10.1016/s1549-3741(04)30071-7]
47. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13(2):138-147 [FREE Full text] [doi: 10.1197/jamia.M1809] [Medline: 16357358]
48. Seidling HM, Schmitt SPW, Bruckner T, Kaltschmidt J, Pruszydlo MG, Senger C, et al. Patient-specific electronic decision support reduces prescription of excessive doses. Qual Saf Health Care 2010 Oct;19(5):e15. [doi: 10.1136/qshc.2009.033175] [Medline: 20427312]
49. Eschmann E, Beeler PE, Schneemann M, Blaser J. Developing strategies for predicting hyperkalemia in potassium-increasing drug-drug interactions. J Am Med Inform Assoc 2017 Jan;24(1):60-66. [doi: 10.1093/jamia/ocw050] [Medline: 27174894]
50. Seidling HM, Al Barmawi A, Kaltschmidt J, Bertsche T, Pruszydlo MG, Haefeli WE. Detection and prevention of prescriptions with excessive doses in electronic prescribing systems. Eur J Clin Pharmacol 2007 Dec;63(12):1185-1192. [doi: 10.1007/s00228-007-0370-9] [Medline: 17786416]
51. Coleman JJ, Nwulu U, Ferner RE. Decision support for sensible dosing in electronic prescribing systems. J Clin Pharm Ther 2012 Aug;37(4):415-419. [doi: 10.1111/j.1365-2710.2011.01310.x] [Medline: 22017267]
52. Topaz M, Goss F, Blumenthal K, Lai K, Seger DL, Slight SP, et al. Towards improved drug allergy alerts: Multidisciplinary expert recommendations. Int J Med Inform 2017 Jan;97:353-355 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.10.006] [Medline: 27729200]
53. Cahill KN, Johns CB, Cui J, Wickner P, Bates DW, Laidlaw TM, et al. Automated identification of an aspirin-exacerbated respiratory disease cohort. J Allergy Clin Immunol 2017 Mar;139(3):819-825 [FREE Full text] [doi: 10.1016/j.jaci.2016.05.048] [Medline: 27567328]
54. Wright A, Hickman TT, McEvoy D, Aaron S, Ai A, Andersen JM, et al. Analysis of clinical decision support system malfunctions: a case series and survey. J Am Med Inform Assoc 2016 Nov;23(6):1068-1076 [FREE Full text] [doi: 10.1093/jamia/ocw005] [Medline: 27026616]
55. van der Sijs H, Mulder A, van Gelder T, Aarts J, Berg M, Vulto A. Drug safety alert generation and overriding in a large Dutch university medical centre. Pharmacoepidemiol Drug Saf 2009 Oct;18(10):941-947. [doi: 10.1002/pds.1800] [Medline: 19579216]
56. Krall MA, Sittig DF. Clinician's assessments of outpatient electronic medical record alert and reminder usability and usefulness requirements. Proc AMIA Symp 2002:400-404 [FREE Full text] [Medline: 12463855]
57. Tilson H, Hines LE, McEvoy G, Weinstein DM, Hansten PD, Matuszewski K, et al. Recommendations for selecting drug-drug interactions for clinical decision support. Am J Health Syst Pharm 2016 Apr 15;73(8):576-585 [FREE Full text] [doi: 10.2146/ajhp150565] [Medline: 27045070]
58. Phansalkar S, Edworthy J, Hellier E, Seger DL, Schedlbauer A, Avery AJ, et al. A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. J Am Med Inform Assoc 2010;17(5):493-501 [FREE Full text] [doi: 10.1136/jamia.2010.005264] [Medline: 20819851]

59. McEvoy DS, Sittig DF, Hickman T, Aaron S, Ai A, Amato M, et al. Variation in high-priority drug-drug interaction alerts across institutions and electronic health records. J Am Med Inform Assoc 2017 Mar 01;24(2):331-338 [FREE Full text] [doi: 10.1093/jamia/ocw114] [Medline: 27570216]
60. Smithburger PL, Kane-Gill SL, Benedict NJ, Falcione BA, Seybert AL. Grading the severity of drug-drug interactions in the intensive care unit: a comparison between clinician assessment and proprietary database severity rankings. Ann Pharmacother 2010 Nov;44(11):1718-1724. [doi: 10.1345/aph.1P377] [Medline: 20959499]
61. Phansalkar S, Desai AA, Bell D, Yoshida E, Doole J, Czochanski M, et al. High-priority drug-drug interactions for use in electronic health records. J Am Med Inform Assoc 2012;19(5):735-743 [FREE Full text] [doi: 10.1136/amiajnl-2011-000612] [Medline: 22539083]
62. Riedmann D, Jung M, Hackl WO, Stühlinger W, van der Sijs H, Ammenwerth E. Development of a context model to prioritize drug safety alerts in CPOE systems. BMC Med Inform Decis Mak 2011 May 25;11:35 [FREE Full text] [doi: 10.1186/1472-6947-11-35] [Medline: 21612623]
63. Seidling HM, Klein U, Schaier M, Czock D, Theile D, Pruszydlo MG, et al. What, if all alerts were specific - estimating the potential impact on drug interaction alert burden. Int J Med Inform 2014 Apr;83(4):285-291. [doi: 10.1016/j.ijmedinf.2013.12.006] [Medline: 24484781]
64. Yang C, Lo Y, Chen R, Liu C. A Clinical Decision Support Engine Based on a National Medication Repository for the Detection of Potential Duplicate Medications: Design and Evaluation. JMIR Med Inform 2018 Jan 19;6(1):e6 [FREE Full text] [doi: 10.2196/medinform.9064] [Medline: 29351893]

Abbreviations

ADE: adverse drug effects
CDSS: clinical decision support system
CPOE: computerized provider order entry
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach, K Fortuna; submitted 26.07.19; peer-reviewed by G Schiff, K Fuji, M Hellaby; comments to author 07.03.20; revised version received 13.03.20; accepted 30.03.20; published 20.07.20.

Please cite as:
Poly TN, Islam M, Yang HC, Li YC(
Appropriateness of Overridden Alerts in Computerized Physician Order Entry: Systematic Review
JMIR Med Inform 2020;8(7):e15653
URL: https://medinform.jmir.org/2020/7/e15653
doi:10.2196/15653
PMID:32706721

Original Paper

Therapeutic Duplication in Taiwan Hospitals for Patients With High Blood Pressure, Sugar, and Lipids: Evaluation With a Mobile Health Mapping Tool

Wei-Chih Kan[1,2], MD; Shu-Chun Kuo[3,4*], MD; Tsair-Wei Chien[5*], MBA; Jui-Chung John Lin[6], DC; Yu-Tsen Yeh[7], BSc; Willy Chou[8,9*], MD; Po-Hsin Chou[10,11*], MD

[1]Department of Nephrology, Chi Mei Medical Center, Tainan, Taiwan
[2]Department of Biological Science and Technology, Chung Hwa University of Medical Technology, Tainan, Taiwan
[3]Department of Ophthalmology, Chi Mei Medical Center, Tainan, Taiwan
[4]Department of Optometry, Chung Hwa University of Medical Technology, Tainan, Taiwan
[5]Medical Research, Chi Mei Medical Center, Tainan, Taiwan
[6]USA Sports Medicine, Sherman Oaks, CA, United States
[7]Medical School, St George's, University of London, London, United Kingdom
[8]Department of Physical Medicine and Rehabilitation, Chiali Chi Mei Hospital, Tainan, Taiwan
[9]Department of Physical Medicine and Rehabilitation, Chung Shan Medical University, Taichung, Taiwan
[10]Department of Orthopedics and Traumatology, Taipei Veterans General Hospital, Taipei, Taiwan
[11]School of Medicine, National Yang-Ming University, Taipei, Taiwan
[*]these authors contributed equally

Corresponding Author:
Po-Hsin Chou, MD
Department of Orthopedics and Traumatology
Taipei Veterans General Hospital
18F, 201, Section 2, Shipai Road, Beitou District
Taipei, 112
Taiwan
Phone: 886 228757557
Email: choupohsin@gmail.com

Abstract

Background: Cardiovascular disease causes approximately half of all deaths in patients with type 2 diabetes. Duplicative prescriptions of medication in patients with high blood pressure (hypertension), high blood sugar (hyperglycemia), and high blood lipids (hyperlipidemia) have attracted substantial attention regarding the abuse of health care resources and to implement preventive measures for such abuse. Duplicative prescriptions may occur by patients receiving redundant medications for the same condition from two or more sources such as doctors, hospitals, and multiple providers, or as a result of the patient's wandering among hospitals.

Objective: We evaluated the degree of duplicative prescriptions in Taiwanese hospitals for outpatients with three types of medications (antihypertension, antihyperglycemia, and antihyperlipidemia), and then used an online dashboard based on mobile health (mHealth) on a map to determine whether the situation has improved in the recent 25 fiscal quarters.

Methods: Data on duplicate prescription rates of drugs for the three conditions were downloaded from the website of Taiwan's National Health Insurance Administration (TNHIA) from the third quarter of 2010 to the third quarter of 2016. Complete data on antihypertension, antihyperglycemia, and antihyperlipidemia prescriptions were obtained from 408, 414, and 359 hospitals, respectively. We used scale quality indicators to assess the attributes of the study data, created a dashboard that can be traced using mHealth, and selected the hospital type with the best performance regarding improvement on duplicate prescriptions for the three types of drugs using the weighted scores on an online dashboard. Kendall coefficient of concordance (W) was used to evaluate whether the performance rankings were unanimous.

Results: The data quality was found to be acceptable and showed good reliability and construct validity. The online dashboard using mHealth on Google Maps allowed for easy and clear interpretation of duplicative prescriptions regarding hospital performance

using multidisciplinary functionalities, and showed significant improvement in the reduction of duplicative prescriptions among all types of hospitals. Medical centers and regional hospitals showed better performance with improvement in the three types of duplicative prescriptions compared with the district hospitals. Kendall W was 0.78, indicating that the performance rankings were not unanimous (Chi square$_2$=4.67, *P*=.10).

Conclusions: This demonstration of a dashboard using mHealth on a map can inspire using the 42 other quality indicators of the TNHIA by hospitals in the future.

(JMIR Med Inform 2020;8(7):e11627) doi:10.2196/11627

KEYWORDS

duplicate medication; mHealth; hypertension; high blood sugar; high blood lipid

Introduction

Cardiovascular disease causes approximately half of all deaths in patients with type 2 diabetes [1,2]. At the population level, an increasing proportion of all cardiovascular events can be attributed to the presence of diabetes [3]. Many epidemiological studies have shown a direct relationship between the levels of blood pressure, glycemia, low-density lipoprotein-cholesterol, and complications of diabetes [4-7]. However, the therapeutic duplication of medication in patients with high blood pressure, high blood sugar, and high blood lipids has attracted substantial attention to prevent the abuse of health care resources.

Duplicative prescriptions refer to situations in which patients receive redundant medications for the same condition from two or more sources [8] such as doctors, hospitals [9,10], multiple providers [11], or as a result of the patient's wandering, in which they move from hospital to hospital for the same condition [12]. Doctor (or hospital) shopping (ie, seeking care from multiple doctors without professional referral for the same or similar conditions) is common in Asia [9,13]. According to Takahashi et al [13], approximately 5.8% of outpatients in Japan self-reported that they visited multiple medical facilities for treatment of the same conditions.

The prevalence of duplicative prescriptions is estimated at 7.4% in Japan [13], which is higher than the rate of 0.43% in Taiwan [14] due to the use of different definitions regarding the Anatomical Therapeutic Chemical (ATC) classification in which the first three five digits are used in Japan and Taiwan, respectively. The management criteria (or tolerance thresholds) of duplicative prescriptions in Taiwan are set at 0.5805%, 0.4273%, 0.5934%, 1.2866%, and 0.9214% for a medical center, regional hospital, local hospital, clinic, and pharmacy, respectively [15], leading the Taiwan National Health Insurance Administration (TNHIA), which operates under the Ministry of Health and Welfare, to strongly express concern about the practice of duplicative prescriptions.

From the perspective of therapeutic safety and excess expenditures, patients who receive medical care from different medical facilities are more likely to receive duplicative prescriptions and suffer adverse drug reactions [9,16-18]. The prevalence of duplicative prescriptions was defined by the TNHIA as the practice of a patient who receives identical medications (based on the first five digits of the ATC) from an identical facility (eg, hospital or clinic) for a period of several overlaid days (ie, total duplicative days/total prescriptive days in a specific period) [14]. A total of 12 indicators of duplicative prescriptions (ie, types of drugs used in the treatment of diseases) have been included and announced quarterly by the TNHIA [19] to help health care providers facilitate management so as to reduce the rate of duplicative prescriptions.

Furthermore, increasing the transparency of hospitals is a requirement to improve administration with regard to patient safety [20-22]; therefore, disclosing the performance of hospitals in effectively controlling duplicative prescriptions to the public is required. If a hospital wants to achieve improvement in patient safety, inspection of a publicly available quality reporting system is essential. Indeed, transparency has been demonstrated as the most powerful driver of health care improvement [23].

By searching for the key words "duplicative prescriptions" on PubMed on April 22, 2020, only one paper [13] was retrieved that reported duplicative prescriptions using social network analysis (SNA). We did not find any study proposing an appropriate method to decrease the number of duplicative prescriptions. That is, when using SNA for interpreting duplicative prescriptions [6], the management perspective is limited in identifying key viewpoints that should be considered in dealing with the duplicative prescription issue.

The SNA approach [24-27] is used to define facilities as the "nodes" of a prescribing network connected to another node (eg, a square box) with a patient duplicative prescription represented as an edge (eg, a connecting arrow). For example, a string of "4 3 1" denotes that node 4 prescribed a duplicative medication via a patient (with a weight of 1) to node 3 using the displayed graphical presentation in which node 4 is connected to node 3 with an arrow.

The objectives of the present study were to (1) assess the attributes of the study data using scale quality indicators, (2) create a dashboard (ie, a control panel on a webpage that collates visual information about an issue or a topic that can be manipulated by readers themselves [28] and can be traced using mobile health [mHealth]), and (3) select the hospital type that shows the best performance in improving duplicate prescriptions of three types of medications (antihypertension, antihyperglycemia, and antihyperlipidemia) using the weighted scores across the types of hospital and performance percentages on an online dashboard. Finally, the Kendall coefficient of concordance (*W*) [29,30] was used to evaluate the unanimity of the performance rankings.

Methods

Study Data

All ratio data for the three types of duplicative prescriptions on the website of TNHIA [19] were downloaded on April 7, 2018 for all registered hospitals in Taiwan. The inclusion criteria were the period from 2010 to 2016 and data recorded in the quarter. Data from a total of 25 quarters (ie, from the third quarter of 2010 to the third quarter of 2016) were included. The exclusion criterion was incomplete ratio data in these 25 quarters. Three types of hospitals, including medical centers, regional hospitals, and district hospitals, were classified and compared. A total of 408, 414, and 359 hospitals were included as study samples for antihypertension, antihyperglycemia, and antihyperlipidemia medications, respectively (Table 1). All data regarding duplicative prescriptions were determined by the ATC classification using the first five digits according to the guideline in Taiwan.

Table 1. Descriptive statistics of hospitals included in the study.

Drug and hospital type	Taipei, n (%)	North, n (%)	Central, n (%)	South, n (%)	Kao-Pin, n (%)	East, n (%)
Antihypertension						
Medical Center (N=20)	7 (35)	2 (10)	4 (20)	3 (15)	3 (15)	1 (5)
Regional Hospital (N=77)	11 (14)	17 (22)	17 (22)	14 (18)	15 (19)	3 (4)
District Hospital (N=305)	29 (10)	62 (20)	86 (28)	40 (13)	76 (25)	12 (4)
Total (N=402)	47 (12)	81 (20)	107 (27)	57 (14)	94 (23)	16 (4)
Antihyperglycemia						
Medical Center (N=20)	7 (35)	2 (10)	4 (20)	3 (15)	3 (15)	1 (5)
Regional Hospital (N=79)	11 (14)	18 (23)	16 (20)	16 (20)	15 (19)	3 (4)
District Hospital (N=308)	29 (9)	63 (20)	88 (29)	47 (15)	69 (22)	12 (4)
Total (N=407)	47 (12)	83 (20)	108 (27)	66 (16)	87 (21)	16 (4)
Antihyperlipidemia						
Medical Center (N=20)	7 (35)	2 (10)	4 (20)	3 (15)	3 (15)	1 (5)
Regional Hospital (N=77)	11 (14)	17 (22)	16 (21)	16 (21)	14 (18)	3 (4)
District Hospital (N=257)	27 (11)	54 (21)	74 (29)	31 (12)	60 (23)	11 (4)
Total (N=354)	45 (13)	73 (21)	94 (26)	50 (14)	77 (22)	15 (4)

Assessing the Quality of Data

Good data quality is necessary to ensure acceptable reliability and validity [31,32].

Therefore, before analysis, the quality of the data was assessed to ensure compliance with responses that may be producible and predictable in similar studies using the following metrics.

Reliability

The reliability (ie, Cronbach α) should be greater than .70 [33].

Dimension Coefficient

The dimension coefficient [34] indicates the strength of unidimensionality, defined as Z/(1+Z), where Z=(a1/a2)/(a2/a3) and the values of a1, a2, and a3 are the eigenvalues of the first three principal components of a scale. The dimension coefficient ranges from 0 to 1; a value greater than 0.67 indicates a unidimensional scale [34].

Convergent Validity

Cronbach α tends to be overestimated. Therefore, it is recommended to rely more on convergent validity (or average variance extracted) and composite reliability values [35] as an assessment of reliability. Convergent validity can be computed as follows:

(**1**)

Where λ is the item loading to the construct domain, λ^2 indicates the communality to the factor, and denotes the measurement error.

Construct Reliability

Construct reliability is also called component reliability or composite reliability, which is expressed by the following formula:

(**2**)

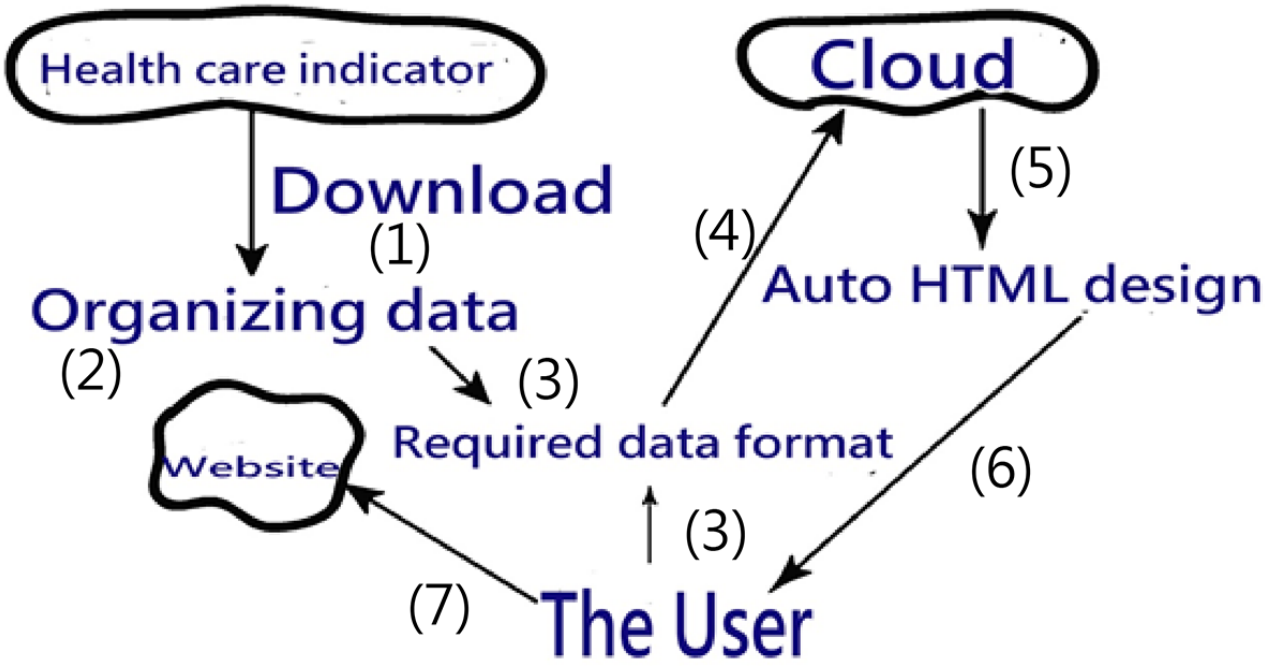
where λ and ε are defined similarly to Equation 1.

Building Online Dashboards on a Map

Figure 1 shows the flowchart of cloud computation to build a quality report card on Google Maps based on quality indicators for data downloaded from the TNHIA website. After organizing the data to fit the required format for uploading, a user can immediately obtain the hypertext markup language (HTML) from the cloud computation through the following three steps: (1) upload data, (2) perform cloud computation, and (3) show

an HTML page that can be downloaded for personal use or public navigation on the website. Interested readers are recommended to view the video demonstrating this process in Multimedia Appendix 1.

Figure 1. Flowchart made on a dashboard. All processes are described in detail in Multimedia Appendix 1.

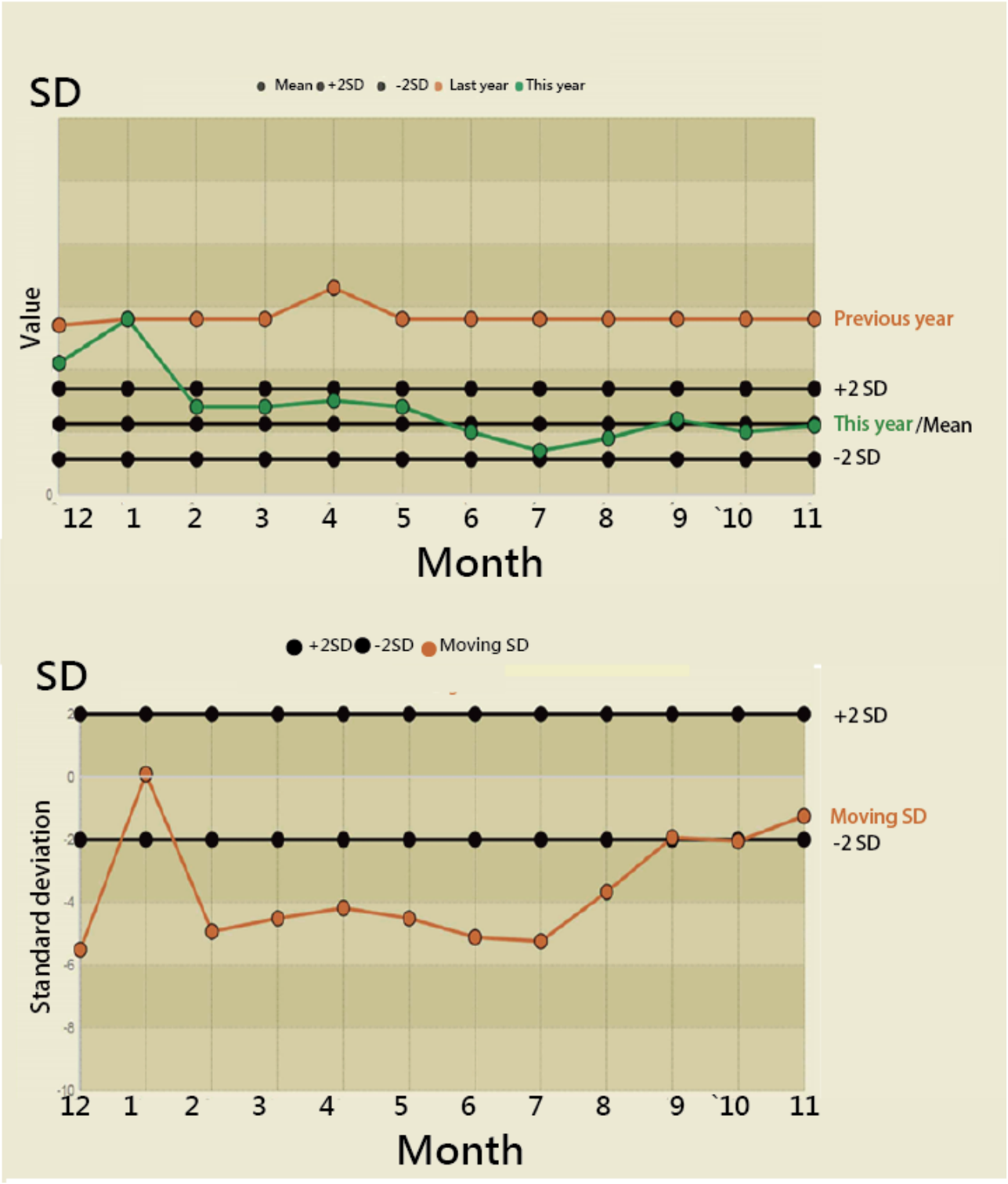


Dashboard Features

The dashboard comprises the following five features: (i) the growth/share matrix of the Boston Consulting Group (BCG) on the map (ie, growth trend on the Y-axis and share on the X-axis) [36,37]; (ii) three traffic light color-coded clusters, which denote the degree of growth/share performance as excellent, fair, and poor; (iii) four quadrants represented by mascots (ie, dogs, question marks or problem children, stars, and cash cows) [37]; (iv) bubbles with a size proportional to product momentum (ie, duplicative prescription ratios in this study); and (v) a control area plotted by the 95% CI (ie, 2 SDs on the two axes).

The growth (on the Y-axis, implying the trend based on recent time points) is determined by the trend via moving the control chart forward to the previous 12 months so that 24 data points yield 12 moving SDs (eg, datasets {–1,–1,–1,–1,–1,–1,–1,–1,–1,–1,–1,1} and {2,2,2,2,2,2,2,2,2,2,2,4} yield an identical correlation coefficient of 0.48 with the time series for 1 to 12), and the share (on the X-axis, indicating the accumulated momentum based on the past) is computed by the mean of the moving SDs (Figure 2 and Multimedia Appendix 2) through which the BCG growth/share matrix can be constructed by the four quadrants on Google Maps (eg, datasets {–1,–1,–1,–1,–1,–1,–1,–1,–1,–1,–1,1} and {2,2,2,2,2,2,2,2,2,2,2,4} yield different momentums of –0.83 and 2.17 across the 12 time points). The study datasets are shown in Multimedia Appendix 3.

Figure 2. Comparison of traditional control chart (top) and moving average control chart (bottom, also see Multimedia Appendix 2) used in this study.



Examples for the Four Quadrants on a Dashboard

The following is a representative algorithm for locating the performance of hospitals on the four quadrants of a dashboard:

- Quadrant I: the dataset {2,2,2,2,2,2,2,2,2,3,4} using the moving control chart forward to the previous 12 months shows continuously increasing growth (ie, y=0.63) with a positive share (ie, x=2.25).
- Quadrant II: the dataset {–1,–1,–1,–1,–1,–1,–1,–1,–1,–1,1,1} shows preparedly increasing growth (ie, y=0.65) with a negative share (ie, x=–0.67).
- Quadrant III: the dataset {–1,–1,–1,–1,–1,–1,–1,–1,–1,–1,–2,–3} shows good performance in controliing duplicative prescriptions with respect to growth (ie, y=–0.63) with a negative share (ie, x=–1.25).

- Quadrant VI: the dataset {2,2,2,2,2,2,2,2,2,2,1,–1} indicates a decrease in growth (ie, y=–0.60) when the share is still positive (ie, x=1.67).

Selecting the Best-Performing Hospital Types in the BCG Growth/Share Matrix

We used the analytic hierarchical process [38] to calculate the weight for each category of performance and then determined the hospital type that performed best in the BCG growth/share matrix according to the following protocol: (i) calculating the percentage in the colorful cluster (ie, the degree of growth/share performance), (ii) multiplying the percentage by the performance weight (ie, 0.5, 0.3, and 0.2 in Figure 3 and the summation equal to 1.0), (iii) summing the weighted score for each hospital type, and (iv) selecting the hospital type that performs best in duplicative prescriptions. The details of the weight calculation are shown in Figure 3.

Figure 3. Calculation of weights for evaluating and ranking hospital performance. In step 1, scores are assigned from 3 (best, green) to 1 (worst, red). In step 2, pair comparison (eg, 3/2=1.5, 2/1=2, 1/3=0.3, etc) is performed to obtain the odds for each cell in the top panel. In step 3, the odds/summation ratio is calculated for each cell in the bottom panel, and the bottom row is averaged to obtain the final weight (eg, 0.5, 0.3, and 0.2).

A. Pair comparisons		Green	Yellow	Red	
		3	2	1	
Green	3	1.0	1.5	3.0	
Yellow	2	0.7	1.0	2.0	
Red	1	0.3	0.5	1.0	
	Summation	2	3	6	
B. Weights		**Green**	**Yellow**	**Red**	**Weight**
		3	2	1	
Green	3	0.5	0.5	0.5	0.5
Yellow	2	0.3	0.3	0.3	0.3
Red	1	0.2	0.2	0.2	0.2
	Summation	1.0	1.0	1.0	

Finally, we used Kendall coefficient of concordance (*W*) [29,30] to evaluate whether the performance rankings were unanimous.

Statistical Analysis

SPSS 19.0 for Windows (SPSS Inc, Chicago, IL, USA) and MedCalc 9.5.0.0 for Windows (MedCalc Software, Mariakerke, Belgium) were used to calculate Cronbach α, dimension coefficients, and other scale quality indicators used in this study. The cloud computation was programmed using the active server pages on the website (see Multimedia Appendix 3). MS Excel Visual Basic for Application (Microsoft Corporation, Redmond, WA, USA) was used to organize the study data.

Results

Data Quality Assessment

The scaling quality for the study data was found to be acceptable (dimension coefficient>0.67 and Cronbach α>.70), indicating that these duplicative prescription ratio data are reliable and consistent with our expectation (Table 2).

Table 2. Quality assessment of the study data.

Type of duplicative prescription	Dimension coefficient	Cronbach α (reliability)	Average variance extracted	Construct reliability
Antihypertension	0.69	.79	0.80	0.99
Antihyperglycemia	0.73	.91	0.85	0.99
Antihyperlipidemia	0.71	.88	0.75	0.98

Building Online Dashboards

The dashboards shown in Figure 4, Figure 5, and Figure 6 show all of the hospitals on the respective maps for duplicative prescriptions of antihypertension, antihyperglycemia, and antihyperlipidemia, in which each hospital is appropriately colored and sized by a bubble. Clicking the bubble shows two kinds of control charts that indicate the traditional 2-year trend and recent 1-year moving average with a trend as illustrated in Figure 2. The control area is divided by the 2 SDs on the X and Y axes, facilitating examining any hospital with extreme performance outside the area. We can also click the icons on the bottom to view the partial type of hospital or the colorful cluster of interest in the left bottom panel. Interested readers may consult references [39-41] or scan the QR codes of the study duplicative prescriptions in Figures 4 to 6.

Figure 4. Dashboard of antihypertension duplicate prescription performance.

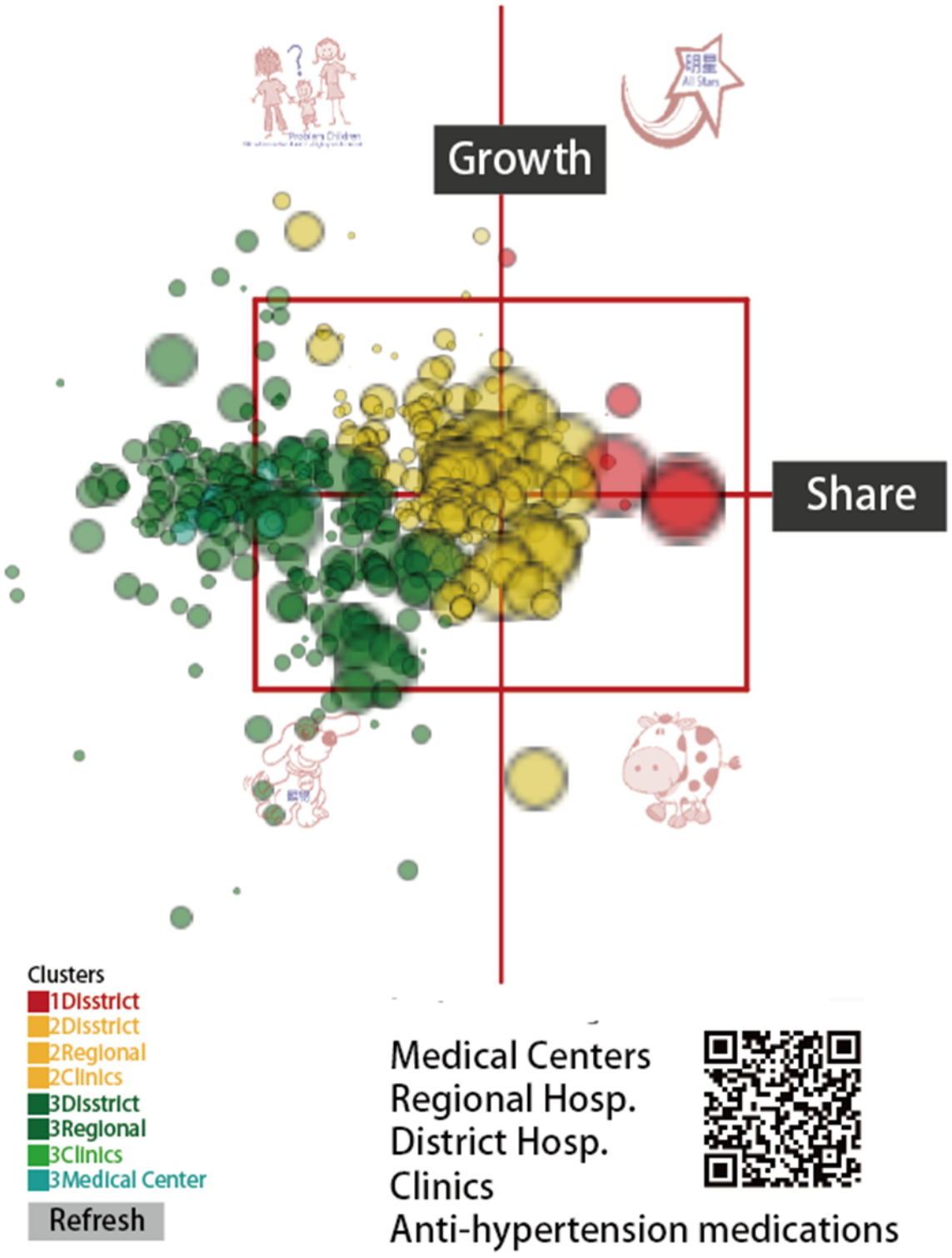


Figure 5. Dashboard of antihyperglycemia duplicate prescription performance.

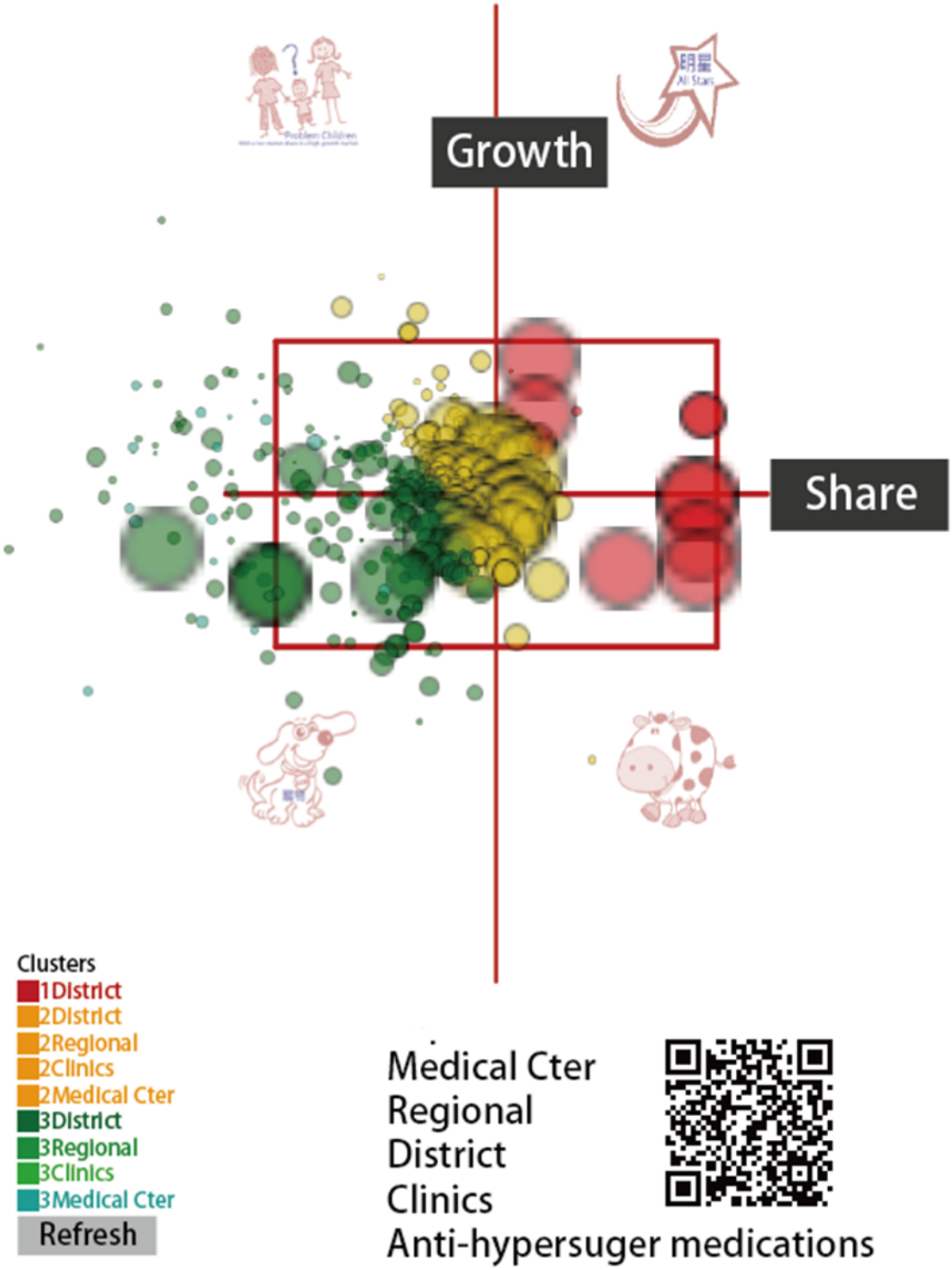
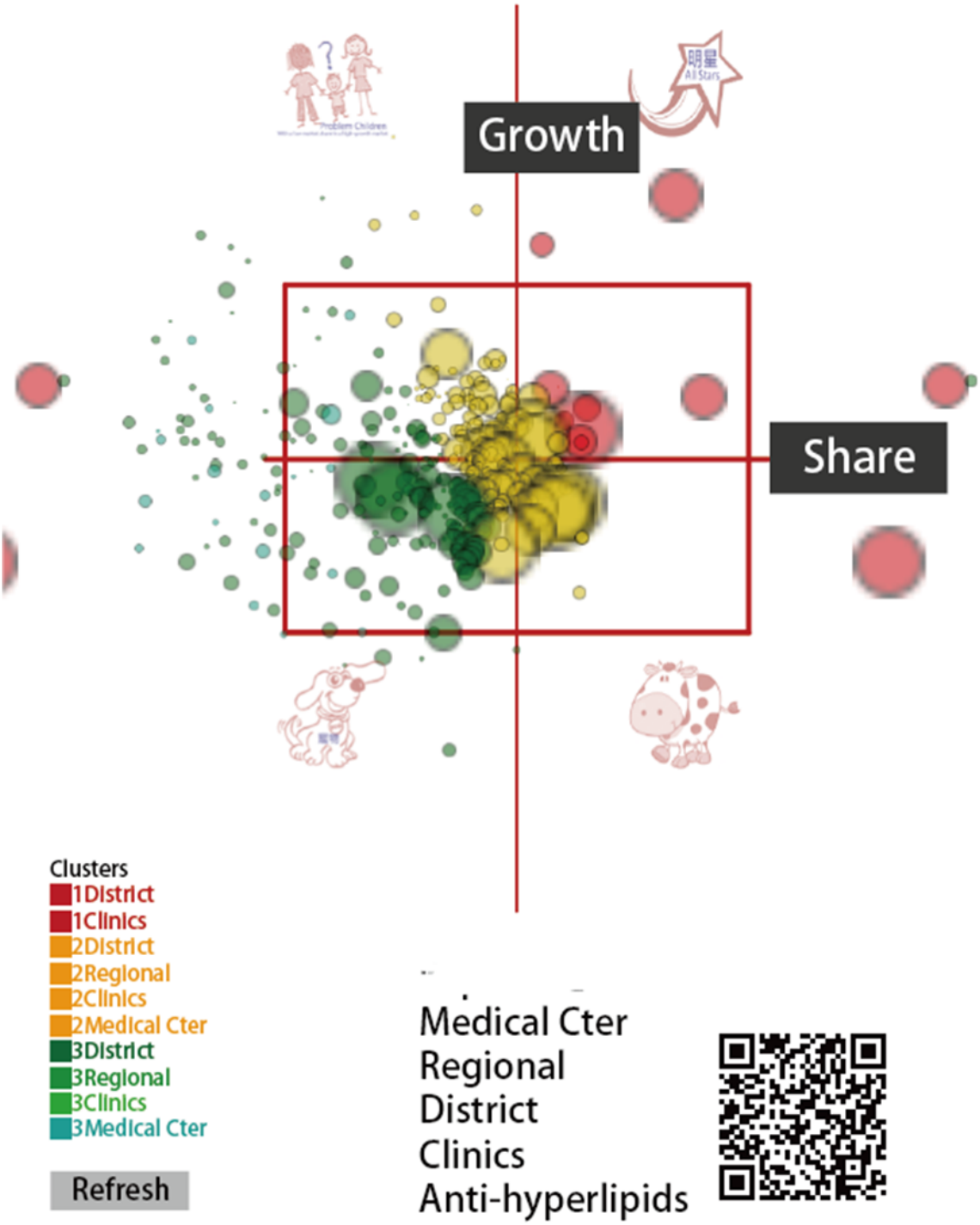


Figure 6. Dashboard of antihyperlipidemia duplicate prescription performance.



Selecting the Best-Performing Hospital Type in Duplicative Prescription Management

As shown in Table 3, the frequency of hospitals in the BCG growth/share matrix on a dashboard showed inconsistent homogeneity among the hospital types, indicating that district hospitals are the largest in number with increasing growth and share (red color code). After summing the weighted scores for each type of hospital in each category of duplicative prescriptions (Table 4), it is clear that medical centers and regional hospital perform best in the growth/share matrix of duplicative prescriptions.

Kendall *W* was 0.781 (χ^2_2=4.67, sum of squares=14, *P*=.10), indicating that the rankings for different types of duplicative prescriptions were consistent (Table 4). Regional hospitals ranked first, demonstrating superiority to the medical centers in the duplicative prescription of antihyperlipidemia medications. Otherwise, Kendall *W* was 1.0 (Chi square$_2$=6.0,*P*=0.05) if the regional hospitals also ranked second.

Table 3. Frequency of the three types of duplicative prescriptions in the four quadrants on the dashboards.

Prescription and hospital type	Red (weight=0.2), n (%)	Yellow (weight=0.3), n (%)	Green (weight=0.5), n (%)	N	Score	Chi square (df=4)	*P* value
Antihypertension,						**64.13**	**<.001**
Medical Center	N/A[a]	N/A	20 (100)	20	50.0		
Regional Hospital	N/A	1 (1)	76 (99)	77	49.8		
District Hospital	8 (2)	170 (56)	127 (42)	305	38.2[b]		
Total	8 (2)	171 (42)	223 (56)	402	N/A		
Antihyperglycemia						**69.91**	**<.001**
Medical Center	N/A	1 (5)	19 (95)	20	49.0		
Regional Hospital	N/A	6 (8)	73 (92)	79	48.4		
District Hospital	13 (4)	156 (51)	139 (45)	308	38.6		
Total	13 (3)	163 (41)	231 (56)	407	N/A		
Antihyperlipidemia						**64.92**	**<.001**
Medical Center	N/A	1 (5)	19 (95)	20	49		
Regional Hospital	N/A	2 (3)	75 (97)	77	49.4		
District Hospital	13 (5)	143 (56)	101 (39)	257	37.3		
Total	13 (4)	146 (41)	195 (55)	354	N/A		

[a]N/A: not applicable.

[b]Score is calculated as: 38.2=(2%×0.2+56%×0.3+42%×0.5)×100.

Table 4. Rankings of hospital type for duplicative prescriptions.

Hospital type	Antihypertension	Antihyperglycemia	Antihyperlipidemia
Medical center	1	1	2
Regional hospital	2	2	1
District hospital	3	3	3

Discussion

Principal Findings

We used dashboards with an mHealth tool to create an animated dashboard that represents the hospital performance sheet of managing duplicative prescriptions in Taiwan. The data quality were acceptable and effectively reflected the reliability and construct validity. The online dashboards enabled easy and clear interpretation of duplicative prescriptions related to hospital performance using multidisciplinary functionalities, demonstrating a trend toward reducing duplicative prescriptions among all types of hospitals. Medical centers and regional hospitals exhibited better performance improvement for reducing duplicative prescriptions for the three types of controlled medications compared with district hospitals. Kendall *W* was 0.78, which indicated that the performance rankings were not unanimous.

Contributions to the Field

Many researchers have published studies based on Google Maps [42-44]. Other studies focused on incorporating the dashboard into a health care report card [45-49], which is worth applying as an informative dashboard to health care settings. However, to our knowledge, this is the first study to build a quality report card as a dashboard, especially using Google Maps, from mHealth.

Making hospitals more transparent [20-22] does not only involve providing a static JPG-format picture but also should include a dynamic dashboard, particularly using a URL to display on mHealth tools for easy comparisons. The dashboards established using the Google Maps application program interface (API) to display health care report cards [46-49] are unique and promising advances in both academic and health care settings for ensuring patient safety against duplicative prescriptions. As such, many other quality-of-care indicators shown on the TNHIA website [50] should be used with an animated dashboard to compare hospital performance rather than traditional static digits or figures [51]. We hope that subsequent studies can report other types of research results using the Google Maps API in the future.

We also found that many district hospitals have incomplete (or missing) data on the ratio of duplicative prescriptions. The reason might be that many district hospitals are significantly affected by the global budget payment system, forcing them to terminate their businesses due to difficult operations in health services.

Management differentiation strategies [52] can be applied through the BCG matrix to review the product portfolio [36,37]. Figures 4 to 6 display the four quadrants derived on market growth (along the Y-axis), relative market share (along the X-axis, indicating the momentum in trend based on previous time points; see Figure 2), and complements of mascots, which are the merits of this study by presenting the BCG matrix with a dashboard on a map.

The use of weights that should sum to 1.0 (as illustrated in Figure 3) differs from the traditional method of performance assessment such as a Likert-type survey using ordinal scores to measure individual performance by summing all item scores with weights not equal to 1.0. We further applied Kendall *W* coefficient to examine whether the performances across all types of hospitals for the three types of drugs were unanimous, demonstrating that the performance rankings were not unanimous and the difference resulted from variation among the drug types.

Implications and Areas for Improvement

Easy Way to Build an Animated Dashboard

Google Maps provides programmers with an API to incorporate coordinates with visual representations and build a dashboard-type report card. We demonstrated the process of creating HTML in the video of Multimedia Appendix 1, which is rarely provided in related research. Interested readers may consult references [39-41] for further details related to Figure 2.

Algorithm for Big Data

The TNHIA website [50] includes many quality-of-care indicators. Intervention is necessary to allow for the systematic collection and analysis of quality-of-care data to assess key quality indicators for all hospitals in a country (or in a region) and provide a "dashboard" feedback to hospitals. The moving control chart is superior to a conventional control chart by providing more valuable information to users. The hospitals with the problem children mascot indicate a readiness to grow. By contrast, the hospitals with the cash cow mascot imply a declining trend. According to the strength of the BCG growth/share matrix, the use of three clusters classified in different colors (red, yellow, and green) and four quadrants are unique and novel in the related literature.

Scale Quality Indicators

As mentioned above, the data quality should be ensured before analysis. This task involves examining the responses that are consistent and reproducible with acceptable reliability and validity [31,32]. Numerous indicators have been proposed to reflect the various ways in which data can be consistent and reproducible. In addition, Cronbach α is a necessary but not a sufficient component of validity [53,54]. Thus, in the present study, we applied other scale quality indicators, including dimension coefficient, average variance extracted, and construct reliability, to examine the quality of the dataset.

Strength of the Study

We evaluated the scale quality with several indicators based on classical test theory. Furthermore, we illustrated the importance of the API in Figure 1 and Multimedia Appendix 4 to demonstrate the infrastructure for applying big data in the cloud computation to build a dashboard-type report card. The BCG matrix incorporated with dashboards can be generalized to many other quality-of-care indicators in the future. The concept of moving control charts [54] can also be applicable and feasible for future use.

Limitations of the Study

Several issues should be considered thoroughly in the future. First, the study data were incomplete, especially for the district hospitals. Thus, inference making, such as for district hospitals with poor performance in controlling duplicative prescriptions, should be conservative. This limitation calls for further research and validation.

Many innovations have been introduced with advances in science and technology, such as the visual dashboard on Google Maps using the coordinates to display and line plots on cloud computation as shown in Figures 4 to 6. However, these achievements are not free of charge. For example, the Google Maps API requires a paid project key for use on the cloud platform, and the line plot also requires payment (to JPowered) for the template used on the website. Thus, the second limitation of the module is that it is not publicly accessible and is difficult to mimic by other authors or programmers for use in a short period of time.

Third, the mascots illustrated in the BCG matrix, such as stars, problem children, cash cows, and dogs, might be inappropriate in health care settings. Other mascots such as Santa Claus, productive cows, or dejected dogs, could refer to appropriate dashboard-type report cards in the future.

Fourth, the scaling quality for the study data was found to be acceptable (ie, dimension coefficient>0.67 and Cronbach α>.70), indicating that these duplicative prescription ratio data are reliable and consistent with our expectation. The dimension coefficients were relatively low (ie, 0.69, 0.71, and 0.73), indicating that all datasets were weak when measuring a one-dimensional feature (ie, duplicative prescriptions). Therefore, there is low confidence when using the result to make an inference for the future. Further studies should pay more attention to the issue of data fitting to the unidimensional requirement.

Fifth, the effect of weights was obvious due to different sample sizes in different hospital types. We normalized the summed weights to be 1.0 and ensured fair comparisons among hospital types across performance categories (ie, red, yellow, and green bubbles). If the percentages of the performance categories differ among hospital types, the weights will affect the assessment results. For this reason, we used an analytic hierarchical process [38] to calculate the weight for each category of performance and then determined the hospital type that performed best in the BCG growth/share matrix, which is worth noting for future assessments.

Conclusion

This study provides a demonstrated platform with an online quality report card on detecting the performance of duplicative

prescriptions to help health care practitioners easily upload data and quickly provide feedback on visual representations on an online dashboard. These dashboards can be used to build an online report card for hospitals under supervision of the public based on mHealth and uHealth in the future.

Authors' Contributions

WC and SC developed the study concept and design, and drafted the manuscript. SC, JU, and YT analyzed and interpreted the data. PH monitored the process of this study. All authors provided critical revisions for important intellectual content. The study was supervised by TW. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
MP3: How to build Google maps for this study.
[TXT File , 0 KB - medinform_v8i7e11627_app1.txt]

Multimedia Appendix 2
The moving average control chart used in this study.
[XLSX File (Microsoft Excel File), 38 KB - medinform_v8i7e11627_app2.xlsx]

Multimedia Appendix 3
Excel dataset.
[XLSX File (Microsoft Excel File), 181 KB - medinform_v8i7e11627_app3.xlsx]

Multimedia Appendix 4
MP3: How to manipulate the mHealth dashboard on Google Map.
[TXT File , 0 KB - medinform_v8i7e11627_app4.txt]

References

1. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HAW. 10-year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med 2008 Oct 09;359(15):1577-1589. [doi: 10.1056/NEJMoa0806470] [Medline: 18784090]
2. Preis SR, Hwang S, Coady S, Pencina MJ, D'Agostino RB, Savage PJ, et al. Trends in all-cause and cardiovascular disease mortality among women and men with and without diabetes mellitus in the Framingham Heart Study, 1950 to 2005. Circulation 2009 Apr 07;119(13):1728-1735 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.108.829176] [Medline: 19307472]
3. Fox CS, Coady S, Sorlie PD, D'Agostino RB, Pencina MJ, Vasan RS, et al. Increasing cardiovascular disease burden due to diabetes mellitus: the Framingham Heart Study. Circulation 2007 Mar 27;115(12):1544-1550. [doi: 10.1161/CIRCULATIONAHA.106.658948] [Medline: 17353438]
4. Emerging Risk Factors Collaboration, Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet 2010 Jun 26;375(9733):2215-2222 [FREE Full text] [doi: 10.1016/S0140-6736(10)60484-9] [Medline: 20609967]
5. Stratton I, Adler A, Neil H, Matthews D, Manley S, Cull C, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ 2000 Aug 12;321(7258):405-412 [FREE Full text] [doi: 10.1136/bmj.321.7258.405] [Medline: 10938048]
6. Stamler J, Vaccaro O, Neaton JD, Wentworth D. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. Diabetes Care 1993 Feb 01;16(2):434-444. [doi: 10.2337/diacare.16.2.434] [Medline: 8432214]
7. Adler A, Stratton I, Neil H, Yudkin J, Matthews D, Cull C, et al. Association of systolic blood pressure with macrovascular and microvascular complications of type 2 diabetes (UKPDS 36): prospective observational study. BMJ 2000 Aug 12;321(7258):412-419 [FREE Full text] [doi: 10.1136/bmj.321.7258.412] [Medline: 10938049]
8. Reeve E, Wiese MD. Benefits of deprescribing on patients' adherence to medications. Int J Clin Pharm 2014 Feb 17;36(1):26-29. [doi: 10.1007/s11096-013-9871-z] [Medline: 24242974]
9. Hsu M, Yeh Y, Chen C, Liu C, Liu C. Online detection of potential duplicate medications and changes of physician behavior for outpatients visiting multiple hospitals using national health insurance smart cards in Taiwan. Int J Med Inform 2011 Mar;80(3):181-189. [doi: 10.1016/j.ijmedinf.2010.11.003] [Medline: 21183402]

10. Worley J, Hall JM. Doctor shopping: a concept analysis. Res Theory Nurs Pract 2012 Jan 01;26(4):262-278. [doi: 10.1891/1541-6577.26.4.262] [Medline: 23556328]
11. Jena AB, Goldman D, Weaver L, Karaca-Mandic P. Opioid prescribing by multiple providers in Medicare: retrospective observational study of insurance claims. BMJ 2014 Feb 19;348(1):g1393-g1393 [FREE Full text] [doi: 10.1136/bmj.g1393] [Medline: 24553363]
12. Pankratz L, Jackson J. Habitually wandering patients. N Engl J Med 1994 Dec 29;331(26):1752-1755. [doi: 10.1056/NEJM199412293312606] [Medline: 7984197]
13. Takahashi Y, Ishizaki T, Nakayama T, Kawachi I. Social network analysis of duplicative prescriptions: One-month analysis of medical facilities in Japan. Health Policy 2016 Mar;120(3):334-341. [doi: 10.1016/j.healthpol.2016.01.020] [Medline: 26876297]
14. Wang W, Wu S, Chien T. Evaluation of therapeutic duplication of medication in patients with high blood pressure, high blood sugar, and high blood lipids between local hospitals, regional hospitals, and medical centers in Taiwan. Journal of Healthcare Management (in Chinese) 2015;16(4):191-205. [doi: 10.2196/11627]
15. Taiwan's National Health Insurance Administration. 2016. Management project of duplicative prescriptions for outpatients in healthcare institutes issued in 2016(Chinese version) URL: https://goo.gl/wM3zRD [accessed 2020-04-22]
16. Bates D, Cullen D, Laird N, Petersen L, Small S, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. JAMA 1995 Jul 05;274(1):29-34. [Medline: 7791255]
17. Tamblyn RM, McLeod PJ, Abrahamowicz M, Laprise R. Do too many cooks spoil the broth? Multiple physician involvement in medical management of elderly patients and potentially inappropriate drug combinations. CMAJ 1996 Apr 15;154(8):1177-1184. [Medline: 8612253]
18. Kinoshita H, Kobayashi Y, Fukuda T. Duplicative medications in patients who visit multiple medical institutions among the insured of a corporate health insurance society in Japan. Health Policy 2008 Jan;85(1):114-123. [doi: 10.1016/j.healthpol.2007.07.003] [Medline: 17728002]
19. Taiwan?s NHIA(. Taiwan's National Health Insurance Administration (TNHIA). 2016. Hospital global budgeting indicators URL: http://www.nhi.gov.tw/AmountInfoWeb/TargetItem.aspx?rtype=2 [accessed 2020-04-22]
20. Callaway E. Toward better administration: treating malignant administrosis with leap-frog feedback. Biol Psychiatry 1994 Jun 01;35(11):827-829. [doi: 10.1016/0006-3223(94)90017-5] [Medline: 8054404]
21. Simpson RL. Improve patient safety by leap(frog)s and bounds. Nurs Manage 2001 Sep;32(9):17-18. [doi: 10.1097/00006247-200109000-00008] [Medline: 17929723]
22. The Leap Frog Group. 2019. Advocating for Transparency URL: http://www.leapfroggroup.org/influencing/advocating-transparency [accessed 2020-04-22]
23. Henke N, Kelsey T, Whately H. Transparency - the most powerful driver of health care improvement? Health International 2011:64-73 [FREE Full text]
24. Landon BE, Keating NL, Barnett ML, Onnela J, Paul S, O'Malley AJ, et al. Variation in patient-sharing networks of physicians across the United States. JAMA 2012 Jul 18;308(3):265-273 [FREE Full text] [doi: 10.1001/jama.2012.7615] [Medline: 22797644]
25. Barnett ML, Christakis NA, O'Malley J, Onnela J, Keating NL, Landon BE. Physician patient-sharing networks and the cost and intensity of care in US hospitals. Med Care 2012 Feb;50(2):152-160 [FREE Full text] [doi: 10.1097/MLR.0b013e31822dcef7] [Medline: 22249922]
26. Landon BE, Onnela J, Keating NL, Barnett ML, Paul S, O'Malley AJ, et al. Using administrative data to identify naturally occurring networks of physicians. Med Care 2013 Aug;51(8):715-721 [FREE Full text] [doi: 10.1097/MLR.0b013e3182977991] [Medline: 23807593]
27. Hanneman R, Riddle M. Introduction to social network methods. Riverside, CA: University of California, Riverside; 2005. URL: https://faculty.ucr.edu/~hanneman/nettext/ [accessed 2020-04-02]
28. Wikipedia. 2019. Definition of dashboard URL: https://en.m.wikipedia.org/wiki/Dashboard_(disambiguation) [accessed 2020-04-03]
29. Kendall MG, Smith BB. The Problem of m Rankings. Annal Math Statist 1939 Sep;10(3):275-287. [doi: 10.1214/aoms/1177732186]
30. Zaiontz C. Real Statistics using Excel. 2019. Kendall's Coefficient of Concordance (W) URL: http://www.real-statistics.com/reliability/kendalls-w/ [accessed 2020-04-22]
31. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol 1993;78(1):98-104. [doi: 10.1037/0021-9010.78.1.98]
32. Green SB, Lissitz RW, Mulaik SA. Limitations of Coefficient Alpha as an Index of Test Unidimensionality1. Educ Psychol Meas 2016 Jul 02;37(4):827-838. [doi: 10.1177/001316447703700403]
33. Lance CE, Butts MM, Michels LC. The Sources of Four Commonly Reported Cutoff Criteria. Organ Res Methods 2016 Jun 29;9(2):202-220. [doi: 10.1177/1094428105284919]

34. Chien T, Shao Y, Jen D. Development of a Microsoft Excel tool for applying a factor retention criterion of a dimension coefficient to a survey on patient safety culture. Health Qual Life Outcomes 2017 Oct 27;15(1):216 [FREE Full text] [doi: 10.1186/s12955-017-0784-8] [Medline: 29078778]
35. Hair JFJ, Hult GTM, Ringle C, Starstedt M. A Primer On Partial Least Squares Structural Equation Modeling (PLS-SEM). Thousand Oaks, CA: Sage; 2020.
36. MacMillan IC, Hambrick DC, Day DL. The Product Portfolio and Profitability--A PIMS-Based Analysis of Industrial-Product Businesses. Acad Manage J 1982 Dec 01;25(4):733-755. [doi: 10.2307/256096]
37. Spee AP, Jarzabkowski P. Strategy tools as boundary objects. Strateg Organ 2009 Apr 15;7(2):223-232. [doi: 10.1177/1476127009102674]
38. Karayalcin II. The analytic hierarchy process: Planning, priority setting, resource allocation. Eur J Operation Res 1982 Jan;9(1):97-98. [doi: 10.1016/0377-2217(82)90022-4]
39. Chien T. Using Google maps to show the performance of duplicate prescription rates in patients with high blood pressure for hospitals in Taiwan. 2018. URL: http://www.healthup.org.tw/kpiall/hyperd.htm [accessed 2020-04-22]
40. Chien T. Using Google maps to show the performance of duplicate prescription rates in patients with high blood sugar for hospitals in Taiwan. 2018. URL: http://www.healthup.org.tw/kpiall/hypers.htm [accessed 2020-04-22]
41. Chien T. Using Google maps to show the performance of duplicate prescription rates in patients with high blood lipids for hospitals in Taiwan. 2018. URL: http://www.healthup.org.tw/kpiall/hyperl.htm [accessed 2020-04-22]
42. Dasgupta S, Vaughan AS, Kramer MR, Sanchez TH, Sullivan PS. Use of a Google Map Tool Embedded in an Internet Survey Instrument: Is it a Valid and Reliable Alternative to Geocoded Address Data? JMIR Res Protoc 2014 Apr 10;3(2):e24 [FREE Full text] [doi: 10.2196/resprot.2946] [Medline: 24726954]
43. Kobayashi S, Fujioka T, Tanaka Y, Inoue M, Niho Y, Miyoshi A. A geographical information system using the Google Map API for guidance to referral hospitals. J Med Syst 2010 Dec 26;34(6):1157-1160. [doi: 10.1007/s10916-009-9335-0] [Medline: 20703591]
44. Kaewpitoon SJ, Rujirakul R, Sangkudloa A, Kaewthani S, Khemplila K, Cherdjirapong K, et al. Distribution of the Population at Risk of Cholangiocarcinoma in Bua Yai District, Nakhon Ratchasima of Thailand Using Google Map. Asian Pac J Cancer Prev 2016 Apr 11;17(3):1433-1436 [FREE Full text] [doi: 10.7314/apjcp.2016.17.3.1433] [Medline: 27039785]
45. Scanlon DP, Shi Y, Bhandari N, Christianson JB. Are healthcare quality "report cards" reaching consumers? Awareness in the chronically ill population. Am J Manag Care 2015 Mar;21(3):236-244 [FREE Full text] [Medline: 25880627]
46. Ivers NM, Barrett J. Using report cards and dashboards to drive quality improvement: lessons learnt and lessons still to learn. BMJ Qual Saf 2018 Jun;27(6):417-420. [doi: 10.1136/bmjqs-2017-007563] [Medline: 29317464]
47. Shepperd S, Charnock D, Gann B. Helping patients access high quality health information. BMJ 1999 Sep 18;319(7212):764-766 [FREE Full text] [doi: 10.1136/bmj.319.7212.764] [Medline: 10488009]
48. Romano PS, Rainwater JA, Antonius D. Grading the graders: how hospitals in California and New York perceive and interpret their report cards. Med Care 1999 Mar;37(3):295-305. [doi: 10.1097/00005650-199903000-00009] [Medline: 10098573]
49. Jaklevic MC. Hospital report-card model in peril. Mod Healthc 1999 Jan 18;29(3):14-15. [Medline: 10345441]
50. Taiwan's National Health Insurance Administration (TNHIA). 2018. Hospital quality-of-care indicators online disclosure on website URL: https://www1.nhi.gov.tw/AmountInfoWeb/TargetItem.aspx?rtype=2 [accessed 2020-04-22]
51. Obstetrics and Gynaecology, University of Toronto. 2018. GTA-OBS Dashboard URL: http://www.obgyn.utoronto.ca/gta-obs-network [accessed 2020-04-22]
52. Porter M. Competitive Strategy: Techniques for Analyzing Industries and Competitors. New York, USA: Free Press; 1980.
53. Feldt L, Brennan R. Reliability. In: Linn RL, editor. Educational Measurement (the American Council On Education/Macmillan Series On Higher Education), 3rd edition. London: Collier Macmillan Publishers; 1993.
54. Chien T, Chou M, Wang W, Tsai L, Lin W. Intraclass reliability for assessing how well Taiwan constrained hospital-provided medical services using statistical process control chart techniques. BMC Med Res Methodol 2012 May 15;12(1):67 [FREE Full text] [doi: 10.1186/1471-2288-12-67] [Medline: 22587736]

Abbreviations

API: application programming interface
ATC: Anatomical Therapeutic Chemical
BCG: Boston Consulting Group
HTML: hypertext markup language
mHealth: mobile health
SNA: social network analysis
TNHIA: Taiwan National Health Insurance Administration

Edited by G Eysenbach; submitted 19.07.18; peer-reviewed by Y Takahashi, K Wright; comments to author 07.01.19; revised version received 06.03.19; accepted 23.03.20; published 27.07.20.

Please cite as:
Kan WC, Kuo SC, Chien TW, Lin JCJ, Yeh YT, Chou W, Chou PH
Therapeutic Duplication in Taiwan Hospitals for Patients With High Blood Pressure, Sugar, and Lipids: Evaluation With a Mobile Health Mapping Tool
JMIR Med Inform 2020;8(7):e11627
URL: https://medinform.jmir.org/2020/7/e11627
doi:10.2196/11627
PMID:32716306

Original Paper

Development and Evaluation of a Smart Contract–Enabled Blockchain System for Home Care Service Innovation: Mixed Methods Study

Shuchih Ernest Chang[1], PhD; YiChian Chen[1], PhD; MingFang Lu[1], MBA; Hueimin Louis Luo[1], MS, MBA

Graduate Institute of Technology Management, National Chung Hsing University, Taichung, Taiwan

Corresponding Author:
YiChian Chen, PhD
Graduate Institute of Technology Management
National Chung Hsing University
145 Xingda Rd
South District
Taichung
Taiwan
Phone: 886 4 22840547
Email: cycx1000@gmail.com

Abstract

Background: In the home care industry, the assignment and tracking of care services are controlled by care centers that are centralized in nature and prone to inefficient information transmission. A lack of trust among the involved parties, information opaqueness, and large manual manipulation result in lower process efficiency.

Objective: This study aimed to explore and demonstrate the application of blockchain and smart contract technologies to innovate/renovate home care services for harvesting the desired blockchain benefits of process transparency, traceability, and interoperability.

Methods: An object-oriented analysis/design combined with a unified modeling language tool was used to construct the architecture of the proposed home care service system. System feasibility was evaluated via an implementation test, and a questionnaire survey was performed to collect opinions from home care service respondents knowledgeable about blockchain and smart contracts.

Results: According to the comparative analysis results, the proposed design outperformed the existing system in terms of traceability, system efficiency, and process automation. Moreover, for the questionnaire survey, the quantitative analysis results showed that the proposed blockchain-based system had significantly (P<.001) higher mean scores (when compared with the existing system) in terms of important factors, including timeliness, workflow efficiency, automatic notifications, insurance functionality, and auditable traceability. In summary, blockchain-based home care service participants will be able to enjoy improved efficiency, better transparency, and higher levels of process automation.

Conclusions: Blockchain and smart contracts can provide valuable benefits to the home care service industry via distributed data management and process automation. The proposed system enhances user experiences by mitigating human intervention and improving service interoperability, transparency/traceability, and real-time response to home care service events. Efforts in exploring and integrating blockchain-based home care services with emerging technologies, such as the internet of things and artificial intelligence, are expected to provide further benefits and therefore are subject to future research.

(JMIR Med Inform 2020;8(7):e15472) doi:10.2196/15472

KEYWORDS

home care service; trust; innovation; blockchain; smart contract; automation

Introduction

Background

Nowadays, human resource management has become a critical issue in the home care industry owing to the advent of an aging society and the growing proportion of double-pay families. The inability to adequately care for elderly people and those with disabilities has created a growing demand for home caregivers. While care centers act as intermediaries in matching home care cases with suitable home caregivers, existing procedures require substantial manual processing (eg, individual case matching, deployment/working status tracking, and employee insurance processing). Consequently, it has become difficult for the existing system to build trust relationships among service providers (care centers), care providers (caregivers), and caretakers [1].

Additionally, majority of the current service platforms store relevant data in their respective local databases, which subsequently leads to information asymmetry, opaqueness, and tampering. These issues have led to a lack of trust among individual systems and have mitigated the level of automation. Moreover, requests must be made to care service providers to retrieve the latest service or insurance status information. Therefore, it is difficult to obtain up-to-date tracking information regarding process flows. The lack of traceability, transparency, and trust among participants increases user concerns (eg, fairness of deployment, insurance status checks, decision-making regarding services provided, and classification/evaluation of care service providers and home caregivers). A practical and radical solution is not only beneficial but also promising for developing a sound home care industry ecosystem [2].

Blockchain and Smart Contracts

Blockchain, as a distributed ledger technology, may solve the aforementioned issues along with its affiliated technology smart contracts. A blockchain can be viewed as a consecutive chain of blocks wherein transaction records are stored. By virtue of its data storage and consensus algorithm, data authenticity and verification are maintained by participating nodes and the distributed network [3,4]. This allows shared duplicates against malicious tampering and results in a trustless operational environment without centralized trusted third parties [5]. Blockchain may provide benefits through innovation, but there may be uncertainty due to technical limitations [6]. However, blockchain allows transparent and auditable transaction records with chronical time stamps, and participants are able to trace related transactions and information flow [7]. Typically, blockchain can help facilitate medical data management [8-11] and drug tracking against potential counterfeit [12].

A smart contract is a computer protocol that can be encoded to digitally facilitate, verify, or enforce the terms or agreement of a contract. Smart contracts have the following characteristics: (1) self-verifying, ability to prove to an arbitrator that a contract has been performed; (2) self-enforcing, enforcement of contractual clauses when predetermined rules/conditions are met; and (3) tamper proof (because of deployment on the blockchain network). Smart contracts may enforce preset rules, implying the potential exemption of trusted third parties. Smart contracts may transform contract terms into programmable logic, with automatic execution if the preset conditions are met [13]. Smart contracts may therefore replace a part of human operation and facilitate more automatic process executions [14]. Ethereum, a popular blockchain platform, supports smart contract execution for various applications by using a Turing-complete language (Solidity) [15,16]. Smart contracts can be automatically executed by programmable codes without manual operation, which, in the long term, can avert human errors and allow workflow automation to achieve higher levels of efficiency [17]. In addition, security is attained because of the tamper-proof characteristics, whereby a more trustful operating system can be established. Certain misconceptions regarding smart contract adoption have been reported [18], including methods regarding the storage/retrieval of data from the blockchain [10,19,20]; however, academic endeavors have not yet discussed how smart contracts can improve process automation in the home care industry.

To facilitate process automation leveraging smart contracts, an event-driven mechanism has been introduced. The event-driven mechanism is a way for the computer system and its affiliated components to manage/handle the required process information flow. It allows applications to communicate with, detect, and react to events. Events can be viewed as a kind of state change. For example, in a home care service scenario, caregivers accept the case, and the state of service matching may change from “undetermined” to “matched.” The occurrence of an event could be further transmitted to applications in the architecture for process managing purposes. A typical pattern is the widely known publish/subscribe or emitter/receiver mechanism. Events are emitted by publishers and received by subscribers who had registered earlier. In this sense, home care event messages are transmitted via functional calls of smart contracts to enable asynchronous communication among system components and computer nodes, and thus, they facilitate procedural manipulation of the overall home care service.

Owing to technical limitations, only data with permanent attributes are considered worthy to be recorded in blockchain because unnecessary storage may increase resource consumption and transaction costs [21]. Generally, existing data required for blockchain transactions may be stored in offline databases, and the interoperability among different systems becomes a critical issue for information exchange [22,23]. The data required by a blockchain system may be retrieved from offline databases by using gateways [24]. For security and privacy concerns, corporations may integrate offline databases with their own developed gateways [25]. In practical implementation, it may be more reasonable to develop self-owned gateways to bridge data access across on- and off-chain systems.

Existing Service Process and Blockchain Roles

Traditional paper-based processes and manual scheduling of service assignments have been heavily adopted in the home care industry. Typical service procedures include service matching, schedule planning, and optimal assignment/dispatch in accordance with critical indicators (available service vacancy, caregiver specialty, and case conditions). Traditional care centers are responsible for manual processing, while specialized

matching systems assist by providing potential matching results. However, the incumbent system may still suffer from poor efficiency and heavy processing for queries and responses.

The major participants of home care services include caretakers, caregivers, and care centers (Figure 1). Each participant has individual siloed databases for data storage. A care center has to confirm and communicate with other units to acquire the latest home care status in that timely updated information is not available from the individual systems of the other two participants. In a short-term employment scenario, a care center may apply for insurance for its caregivers to cope with emergencies or unexpected circumstances. Insurance companies have been reluctant to provide relevant short-term products owing to the complexity of the application/cancellation process and costs related to manual processing. While accidental events may be inevitable during service delivery, caregivers' employment and demands from care recipients may be influenced by the lack of insurance products in the home care industry.

Owing to blockchain data storage and consensus algorithms, home care service participants may benefit from a common shared ledger and enjoy process automation facilitated by smart contracts. Figure 2 illustrates the roles that blockchain and smart contracts could play in improving the current issues faced by the home care industry in terms of traceability, timeliness, interoperability, and cost. The proposed design may reduce the degree of human intervention, thereby increasing the integrity and performance of the system during operation. Relevant smart contracts are designed to connect the entire system's process flow, enabling the automation of service processes. Moreover, parties may perceive a higher level of freedom to request status information, which in turn enhances transparency and traceability during the service process. Additionally, the event mechanism of smart contracts can instantaneously disseminate information to all parties and enable response to important events.

Figure 1. Overview of the current matching system.

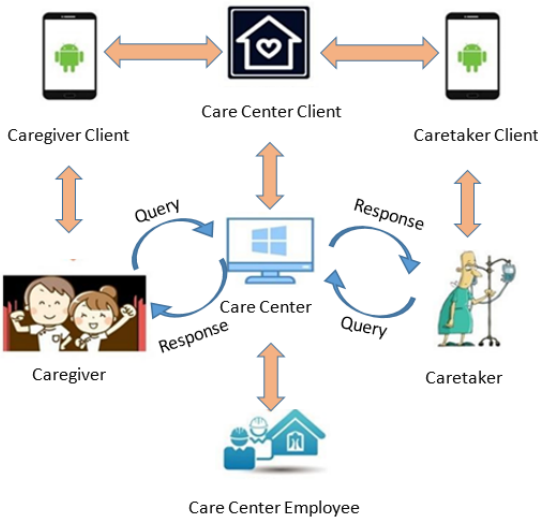
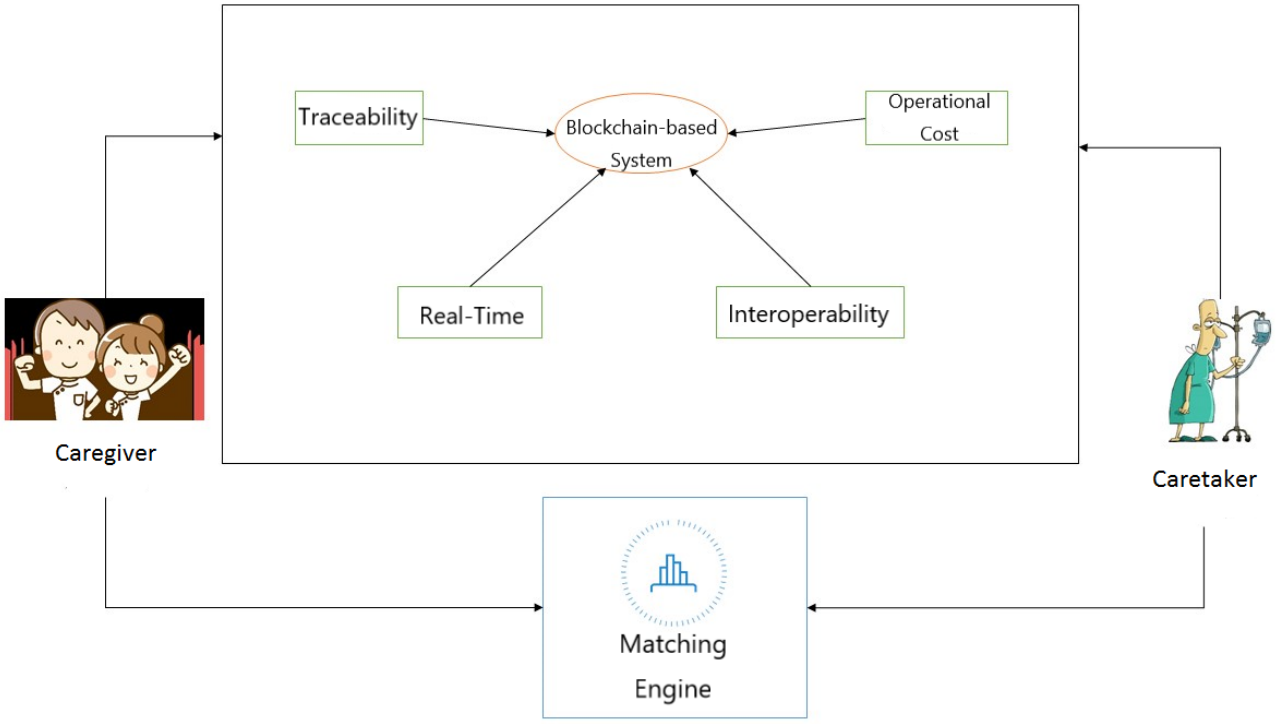


Figure 2. The role of blockchain in improving the home care service system.



Study Objective

A potential blockchain-based solution (ie, a blockchain and smart contract–enabled home care service system) is proposed to resolve the aforementioned issues affecting the home care industry. We not only developed a step-by-step system design protocol but also evaluated the core functions and the proposed automated processes to test the system's feasibility.

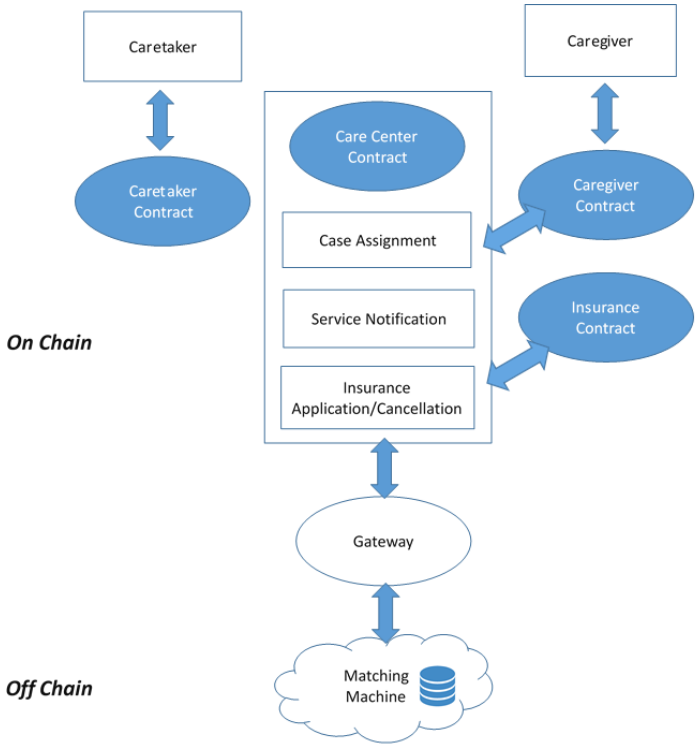
Methods

Overview of the Blockchain and Smart Contract–Enabled Home Care System

To address the major issues in delivering home care service in terms of case assignment, information notification, and insurance procedures, our research team consisting of the authors and two research assistants (programmers) worked part-time and developed a blockchain-based home care service system. The study took place from October 2017 to January 2019. From a process re-engineering perspective, we combined the unique features of blockchain and smart contracts in the design of an integrated system to solve major pain points using an existing home care system. Figure 3 shows the proposed framework of the blockchain-based home care service system. The system's operation begins by receiving matching cases and enabling the automatic deployment of caregivers. A caretaker receives automatic notifications after a caregiver's acceptance. The caretaker may either decide to start the home care service or not, and the system may automatically enroll/cancel the insurance for the caregiver. A potential solution is designed and explored to achieve greater process efficiency and timeliness in terms of caregiver matching, task assignment, service notification, and insurance-related processes. The proposed design allows improved transparency/visibility on process status transitions by introducing event-driven smart contracts on a blockchain-based platform. Additionally, preset criteria for insurance claims are designed to activate the claims process once triggering conditions are met.

For the above-described purposes, this study focused on the interactions among care centers, caregivers, caretakers, and insurers, and thereby designed the following four kinds of smart contracts: care center contract, caregiver contract, caretaker contract, and insurance contract. The proposed system integrates with the existing matching engine, which provides optimal matches, and retrieves off-chain matching results via a gateway for further use in the care center contract. This contract deals with task assignment, service notifications to caregivers/takers, and insurance-related processes.

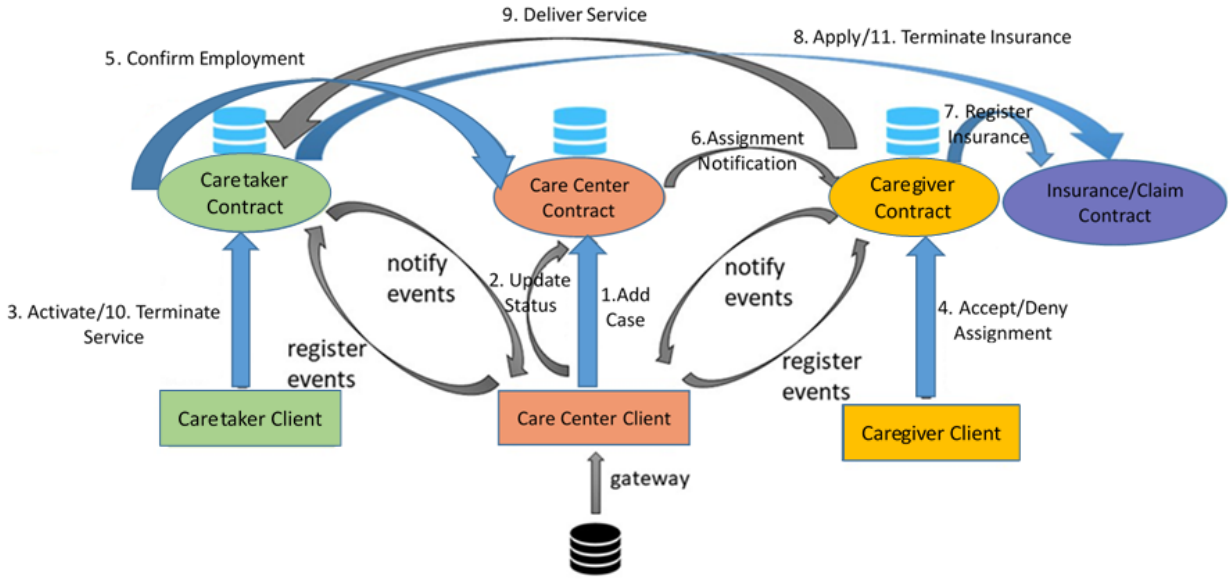
Figure 3. Overview of the blockchain and smart contract–enabled home care system.



Process Flows and Business Model of the Proposed System

Figure 4 presents the framework for the proposed system to elucidate the design of smart contract interactions from a project management perspective. Home care service participants may track the system's status in a timely manner and be informed of data relevant for decision-making via event-driven notifications. Smart contracts were further utilized to design an insurance claims procedure that may simplify the complex manual review processes and enable claims procedures to be automatically executed.

Figure 4. Framework of the proposed blockchain-based home care system.



System Process Flow

Initially, an external matching engine uploads service-matching results and related information to the care center contract through a gateway via a standard information interface provided by the system. Thereafter, care center employees dispatch task assignments to the caregiver contract. Automatic assignment notifications are then forwarded to the caregiver for his/her final decision. The caregiver contract receives the caregiver's response and then delivers a confirmation message (regarding accept/reject results) to the care center contract. Once an assignment is confirmed, the care center contract executes the

notification function previously coded in the caretaker contract and informs the caretaker of related service information. The caretaker may decide to accept/reject the service. If the caregiver accepts, he/she makes a request to activate the home care service. Meanwhile, the care center contract creates an insurance application for the caregiver. When the home care service is completed, the caretaker can ask the care center contract to terminate the home care service.

This system utilizes smart contracts to connect holistic service processes. The event-driven mechanism allows decision information to be transmitted to the relevant decision makers. Once decisions are made, functional linkage would be redirected to the original service processes. Despite the decision points, all service processes in this system are executed by corresponding smart contracts without manual intervention.

New Business Model

The proposed system can serve as a public platform for improved service status tracking and as a potential solution for caregiver insurance issues.

Blockchain-Based Platform

The existing home care service system adopts a centralized model for operational management. This study presents a new type of business model for participating stakeholders. In the current home care system, an individual caregiver/taker makes requests to either public sector or social service units for case matching and passively waits for vague matching results. Thereafter, authorized care centers take charge of care service processing till the end of the service. This existing model suffers from a lack of transparency and poor efficiency owing to its centralized operation, making it difficult to establish a fair and impartial evaluation mechanism.

The proposed system seeks to provide a solution that simultaneously favors both the caregiver/taker and care centers. First, caretakers may not be equivalent in their economic capabilities. Existing social service providers, religious/charity parties, and public sector organizations have been undertaking related projects to meet the increasing demand from stakeholders. The proposed system allows for the inclusion of home care services in these projects by providing a new operational model for developing a fair and open matching platform. Apart from professional caregivers, volunteer caregivers can also be incorporated to extend the home care service capability to support policy. In these projects, it is technically feasible to add new categories of caregivers in project implementation without incurring too much overload in terms of computing algorithms and system design. This operational model can be designed by the integration of on- and off-chain databases. While the matching engine enables massive data processing by using complex algorithms, computational processing may be assigned to the offline system to generate the final matching results for upload to the on-chain system.

The proposed system can be treated as a service execution engine that is capable of (1) tracking the status of caregivers' assignment and execution process, (2) automatically notifying caretakers about service commencement and dominant rights regarding service initiation/termination, and (3) introducing newly developed insurance applications for improving caregivers' personal protection.

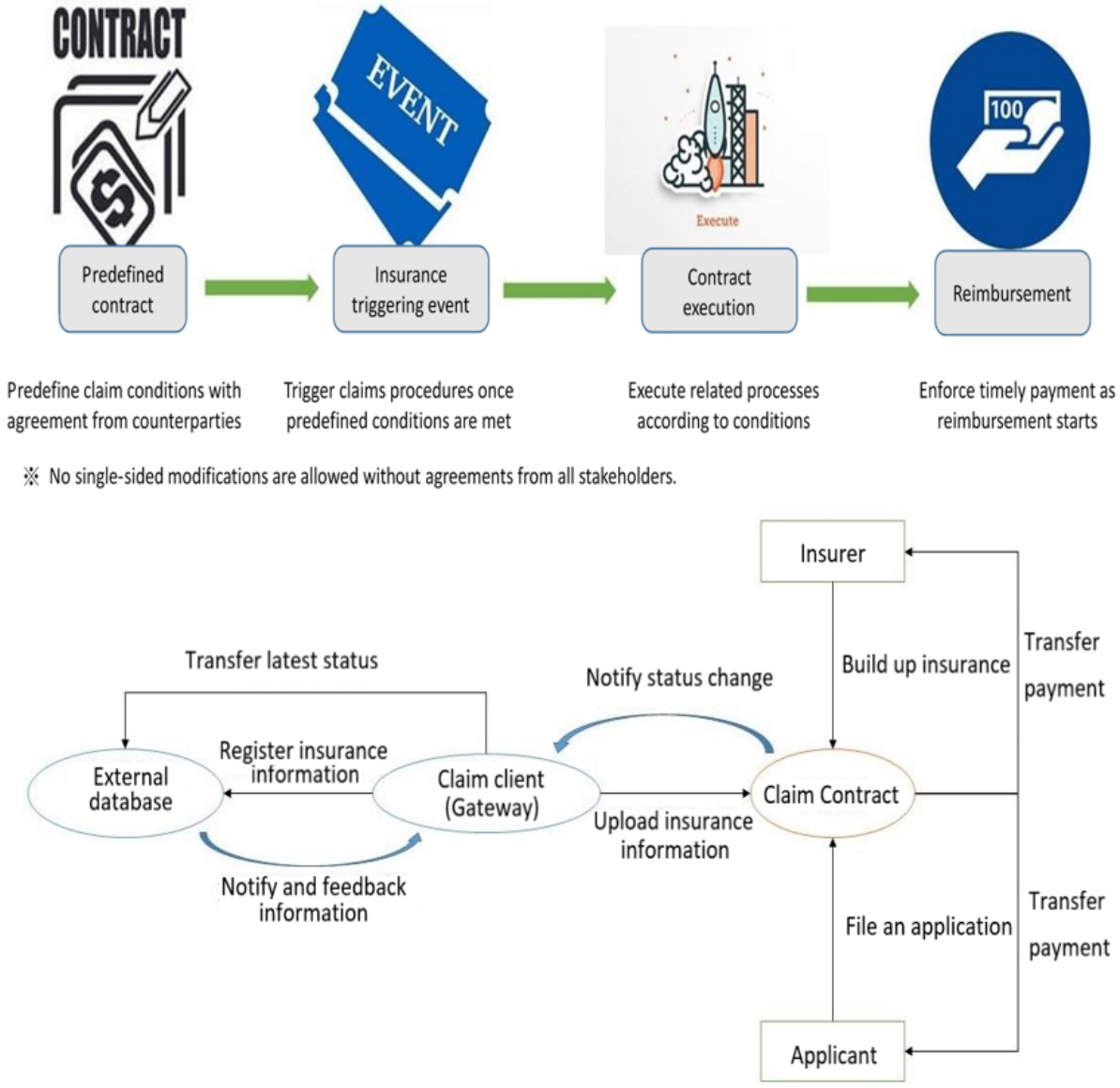
New Insurance Products Based on the Blockchain System

Traditional insurance claims procedures may take weeks or even months for reimbursement owing to the massive volume of manual processing involved. Increased administrative costs may lead to a higher premium for applicants. A smart contract–enabled insurance process transforms enforcement terms and rights and obligations into programmable codes. Once an insurance event has occurred, such as a traffic accident, corresponding information can be uploaded to the claims contract, which in turn makes a comparison with preset claims conditions. If these conditions are met, claims proof-related information is recorded on the blockchain and automatic claims procedures are activated. The reimbursement is automatically executed without any manual intervention (Figure 5). In this regard, smart contract–enabled insurance may not only reduce considerable administrative costs/time but also provide transparent disclosure on claimed items and criteria. The overall claims procedures are open to all stakeholders and are enforced according to contract terms. Such an insurance application may develop better trust among stakeholders, thereby ensuring their rights and obligations.

In practice, the insurance company automatically files the corresponding insurance policy whenever a home care case is confirmed and established by the care center contract according to a comprehensive evaluation of caregiver/caretaker background and other risk considerations during service delivery. The subsequent process is enabled by an insurance smart contract accordingly, and claims for the caregiver are conducted whenever insurance conditions are met with status changes/updates.

This study explored a new insurance product wherein no particular insurer is necessarily required to act as an insurance carrier since the main design rationale is to set methods of developing insurance policies and filing insurance applications as payable functions. The use case of the claims contract is generated by insurers, which enables policy reserves to be deposited in the contract. When applicants file insurance applications, the policy reserves are deposited in the claims contract. Once a claim's conditions are met, the proposed system automatically executes the settlement of loss reserves from the total policy reserves according to the claims ratio between applicants and insurers. This allows for the exemption of manual review/auditing procedures with automatic claims enforcement in place against intended noncompliance.

Figure 5. Framework of the smart contract–enabled insurance claims procedure.

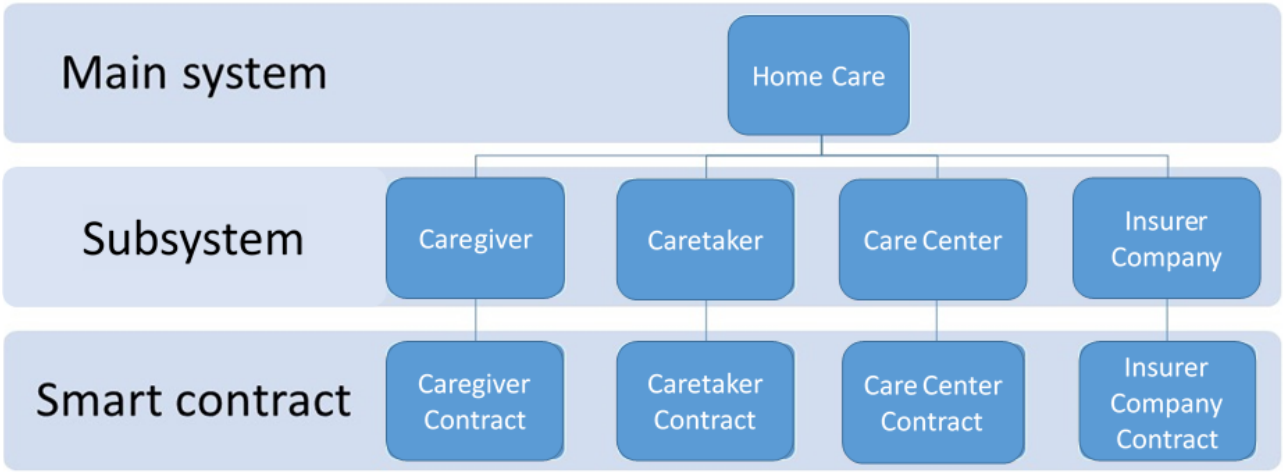


Object-Oriented Methodology and Unified Modeling Language Analysis

This study uses object-oriented application/design methodologies for the system analysis and design [26,27]. Unified modeling language diagrams were utilized for system modeling [28]. In unified modeling language, use case diagrams are used to represent system functions and the interactions between the actors and these functions. In a static configuration, class diagrams are used to describe the data structures of smart contracts and the methods by which users could access their services. In a dynamic configuration, sequence diagrams are used to depict how actors interact with smart contracts to fulfil compulsory functions and to specify the operational processes of the system. More detailed steps of system design are provided in Multimedia Appendix 1. The home care system comprises four subsystems with affiliated smart contracts (ie, caregiver, caretaker, care center, and insurer contracts) (Figure 6). The proposed framework enables short-term home care service assignment, automatic process state notifications to stakeholders, and insurance application/cancellation by incorporating the insurers' contracts.

To better present the feasibility of the proposed system, we conducted an implementation test to demonstrate the functionality of each system component and operational process. More detailed test guidelines, strategies, and procedures are provided in Multimedia Appendix 2.

Figure 6. Architecture of the home care service system.



Implementation and Validation

Implementation Platform

This study's implementation aimed to validate the feasibility of the smart contract–enabled home care service system. Since front-end platforms require further study, the development of the blockchain system, including the validation of interoperability and status transitions among smart contracts, has become more important. The Chrome browser (Google) and Remix software with a JavaScript VM environment were required for system validation. We utilized Solidity for smart contract programming and Solidity compiler 0.4.18 for the compilation of Solidity contracts that were deployed in the blockchain system for further testing [29].

Implementation Procedures

The following assumptions were made before implementation to simplify the procedures of feasibility validation: (1) For each caretaker, only one care service is accepted by a caregiver on a daily basis; (2) A caregiver only provides service to one caretaker; and (3) A care center provides only one insurance application per caretaker and caregiver.

Comparative Analysis of the Existing and Proposed Systems

A step-by-step comparison of care service workflow was performed to point out differences between the existing and proposed systems. In particular, the examination in terms of critical relevant dimensions (including transparency, traceability, level of automation, counterfeit/fraud proofing, and management of insurance/welfare) may help shed light on the competitive advantages to be availed of when adopting a blockchain-based system. The blockchain and smart contract features are major criteria for giving qualitative measures on individual steps. Through step-by-step workflow comparisons, valuable information of system comparisons was extracted to provide overall insights from a qualitative perspective.

Questionnaire Survey

In this study, we also administered a questionnaire survey to investigate the feasibility of a blockchain-based system by conducting comparisons with the existing system. The selected respondents were participants in a final term exhibition of the Digi+ Talent Accelerator & Jumpstart Program, administered by the Industrial Development Bureau, Minister of Economics Affairs, Taiwan. This incubator program is a 4-year program, and the first-year program was completed during the period from July 1 to December 31, 2017. The exhibition gathered Fintech scholars, engineers, and practitioners in blockchain-related fields from academia, industry, and public sectors in Taiwan. Participants were selected and filtered during enrollment to make sure that they were qualified with matching and had appropriate knowledge and experience, particularly related to blockchain and home care services. The questionnaire was validated by an expert panel invited from the selected respondents, while data were collected in December 2017 and analyzed in early 2018. The surveyed aspects of the questionnaire were divided into the following five major constructs: the care center, caregiver, caretaker, insurer, and public sector. The four items of traceability, efficiency, automation level, and management were considered critical for measuring the feasibility of the proposed system. Purposive sampling was adopted for scholars, experts, practitioners, and graduate students, and a total of 50 home care service respondents were collected in a 7-day period. The selected respondents were considered suitable for the questionnaire survey as they were familiar with home care services and were knowledgeable about blockchain technology and smart contracts.

Results

Results From the Implementation Test

For every use case scenario, we developed class codes to implement needed service flow as described in the previous section and also detailed in Multimedia Appendix 1. Actually, we conducted all necessary functionality tests to ensure that the implemented service flow functions smoothly as expected by design. The service flow is home care case establishment, task assignment, service notification, service activation, automatic insurance application, service termination, and insurance cancellation. While testing all service flow sequences, we extracted related testing screenshots, which are shown in this paper and in Multimedia Appendix 2, to demonstrate that interactions among smart contracts correctly match the designed

service flow scenarios. In this paper, we only present two test result–related screenshots (Figures 7 and 8) to demonstrate the real interactions among smart contracts with service status changes. More comprehensive test result–related screenshots together with their matching service flow scenarios are presented in Multimedia Appendix 2.

Figure 7. Illustration of the caretaker contract status after task assignment.

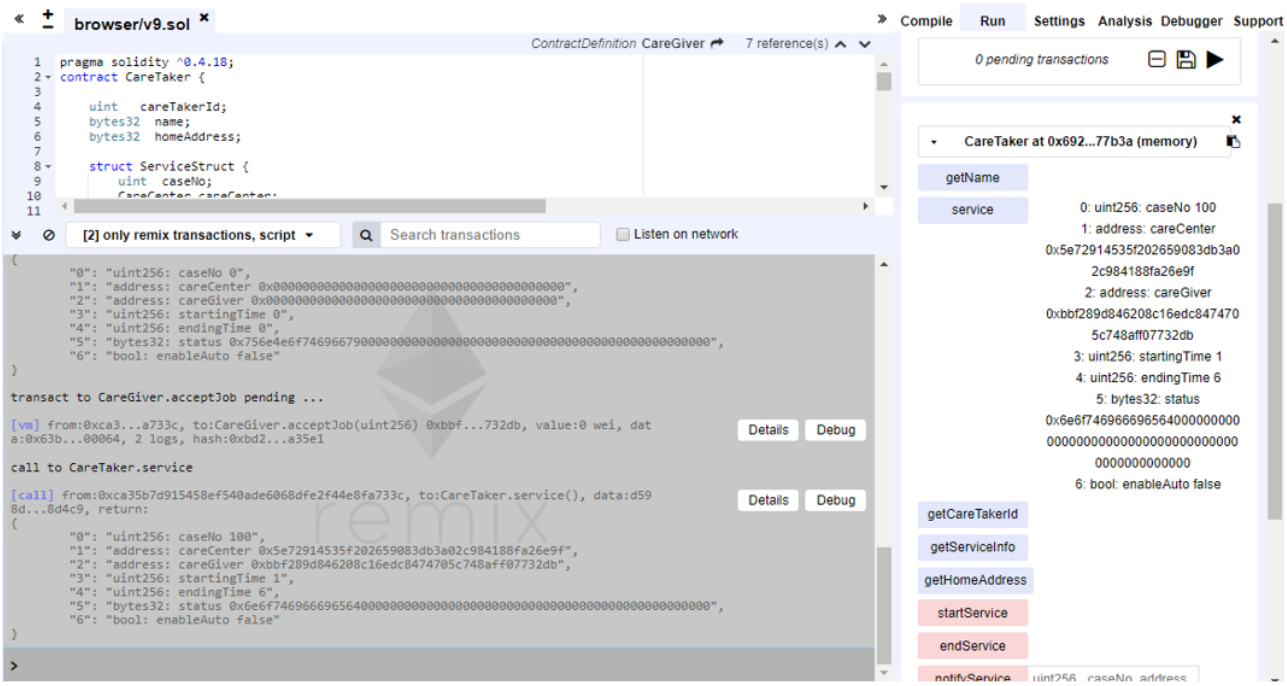
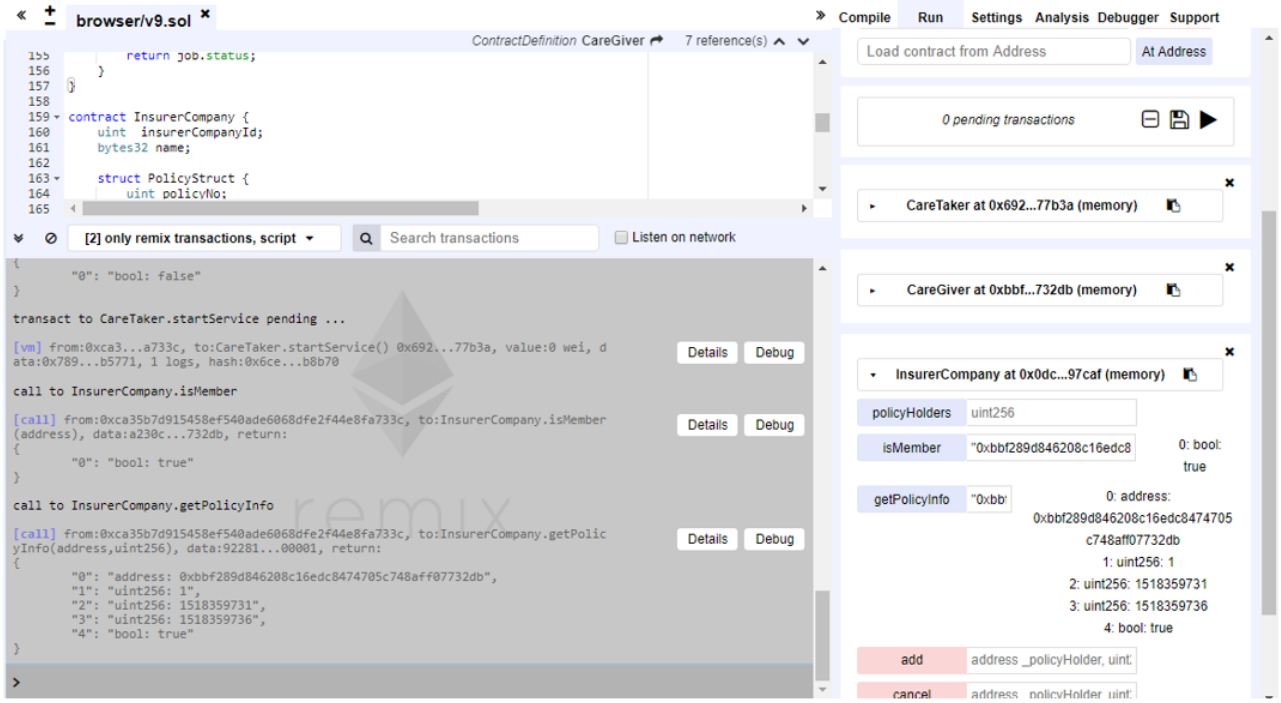


Figure 7 illustrates the caretaker contract status after task assignment. The care center automatically notifies the caretaker of related service information upon receiving caregiver acceptance. Again, by using the service function, related service information is available in the bottom corner of the right-hand side column. For example, case number 100 is reported with care center and caregiver address details. Figure 8 presents the notification and status of the insurance policy after service activation. The system is designed in a manner such that the care center may automatically file an insurance application for the caregiver once the care service is activated. Users could inquire about the insurance-related status by using the getPolicyInfo function to get insurance policy details.

Figure 8. Illustration of the insurance application status after service activation.



Illustrations From Comparative Analysis

To have a better understanding of how blockchain and smart contracts enable improving performance against the existing system, we conducted a step-by-step comparison of the care service workflow for reference (Multimedia Appendix 3). In addition, we compared the proposed system with the existing system in terms of five major constructs (Table 1). The major differences were reduced administrative costs, increased automation, and timely tracking of system status.

Table 1. Comparison of constructs between the existing system and the proposed system.

Construct	Existing system	Blockchain-based system
Transparency	1. Highly centralized governance structure in single home care authority	1. Smart contract-enabled management and review of aggregated information and evaluation indices
	2. Lack of trust due to unavailability of monitoring	2. Information transparency with public-access authority
	3. Discrepancies in auditing/evaluation	3. Improved trustworthiness of the overall system
Counterfeit	1. Manual interference during operations	1. Replacement of manual operations by smart contracts
	2. Difficulties in waiving artificial alterations and possibility of tampering	2. Mitigation of alterations and tampering of previous recorded data due to permitted access control
Traceability	1. Job vacancies by carer queries	Automatic notification by an event-driven mechanism
	2. Manual notification to the case	
	3. Update latency on the latest status	
Level of automation	Manual processing on case assignment, notifications, and insurance-related procedures	Smart contract–enabled process automation
Insurance and welfare	1. Complexity regarding the application/cancellation of existing products	1. Newly developed insurance product to facilitate the review process in terms of claims proportion
	2. Lack of insurance coverage in the home care industry	2. Automatic claiming process and imbursement enabled by preset terms or agreement
		3. Large reductions in administrative costs and better insurance offers for applicants

Findings From the Questionnaire Survey

Since the existing and proposed systems are not completely independent of each other, this study utilized a dependent *t* test to analyze the mean differences. Given that the sample was normally distributed and possessed homogeneity of variance, this study conducted the analysis using a dependent *t* test. Table 2 reports the means and SDs for the existing and proposed blockchain systems in terms of the five major constructs. Further data analysis showed that significant differences existed ($P<.001$) and each function in the proposed blockchain system was superior to that in the existing system. The proposed blockchain system was therefore validated for better performance and excellence than the traditional system. More detailed information about the content of the questionnaire is provided in Multimedia Appendix 4.

This study analyzed the potential reasons for the most significant items in terms of their *t* values. For care centers, the respondents generally considered that the proposed system may outperform the existing system by providing better traceability of caregiver assignments and service status. For caregivers, the automatic notifications from the blockchain system provide them with automatic service status updates and notifications along with the subsequent activation of claims procedures. For caretakers, respondents think that improved efficiency of insurance claims procedures can be achieved since the system enables automatic insurance application/cancellation. For insurers, the interactive processes in the smart contract design allow companies to automatically execute short-term insurance application/cancellation and facilitate claims procedures. Finally, the public sector may benefit from blockchain's immutable transaction records for easy tracking, case matching/assignment, and related insurance information.

Table 2. Comparison of construct scores between the existing system and the proposed blockchain-based system.

Item (function)	Existing system score[a], mean (SD)	Blockchain system score[a], mean (SD)	*t* value[c]
Care center (beneficial to…)			
Traceability[b]	3.28 (0.991)	4.20 (0.639)	5.619
Efficiency[b]	3.28 (1.070)	4.10 (0.763)	4.499
Automation[b]	3.14 (1.088)	4.14 (0.857)	4.763
Management[b]	3.18 (1.024)	4.08 (0.724)	5.161
Caregiver (beneficial to…)			
Traceability[b]	3.24 (1.041)	4.28 (0.730)	6.340
Efficiency[b]	3.24 (1.061)	4.14 (0.783)	5.306
Automation[b]	3.10 (1.035)	4.26 (0.853)	7.029
Caretaker (beneficial to…)			
Traceability[b]	3.12 (1.100)	4.12 (0.849)	5.052
Efficiency[b]	3.00 (0.990)	3.96 (0.903)	6.354
Insurer (beneficial to…)			
Traceability[b]	3.18 (1.063)	4.10 (0.678)	5.039
Efficiency[b]	3.10 (1.055)	4.18 (0.691)	6.404
Automation[b]	2.96 (1.009)	4.22 (0.648)	7.584
Management[b]	3.16 (0.976)	4.22 (0.790)	5.986
Public sector (beneficial to…)			
Traceability[b]	3.34 (0.939)	4.30 (0.647)	6.354
Efficiency[b]	3.26 (1.065)	4.32 (0.621)	6.066

[a]Score ranging from 1 (strongly disagree) to 5 (extremely agree).

[b]$P<.001$.

[c]Degree of freedom df=49.

Discussion

Principal Findings

The proposed system contributes to the understanding of blockchain-based application in the home care service industry. The aforementioned system design, implementation, and testing showed that the system achieved the expected requirements in terms of delivering the needed home care service functionalities and harvesting the desired blockchain benefits. In the real world, when a critical change in service status occurs, related events are emitted to notify registered clients. These functional demands are actually achieved while the design of nonsynchronization may cope with functional circumstances in a real system. The structural design, development, implementation, and feasibility testing of the proposed blockchain-based system reflect its potential for home care service applications. The step-by-step design protocol also provides a reference for future academic research. However, for the change from prototype to real implementation in an actual home care scenario, major challenges and implementation issues need to be addressed and discussed to shed light on orientations in a future study.

Practical Implementation Issues

Generally, the major challenges to blockchain implementation in the real world on a large scale going forward can be classified into technical, organizational, and environmental aspects [30], and we need to address such technical, organizational, and environmental issues to make our blockchain-based home care service actually work in real life. System efficiency is an extremely important technical challenge to address because the system design will not be practical unless the efficiency of our blockchain-based system is acceptable [31]. In terms of efficiency improvement, researchers suggested exploring various technological configurations on blockchain platforms, consensus protocols, data exchange mechanisms between on-chain and off-chain systems, block sizes, etc [31]. Cultural shift and mindset are also critical factors when organizations attempt to embrace blockchain applications. From paper-based procedures to digitization processes, high resistance from transformation needs to be either mitigated by employee training or ameliorated by more incentives and perceived benefits. In addition, the

unwillingness for data disclosure among distributed parties may hinder large scale information sharing, thus affecting the overall performance of blockchain-enabled systems. For example, care providers may be reluctant to share data with insurers owing to the lack of concrete incentives and legacy system users may have security/privacy concerns with regard to adopting new technology. Finally, government policies or regulatory support for emerging blockchain technologies could be influential. The implementation processes for health care administrations could be affected by such environmental factors. Without proven credibility from leading pilot projects, large-scale adoption may not be viable and promising for various applicable industries to harvest desired blockchain benefits.

Specifically, to accelerate settlement and claims procedures for our blockchain-based home care service, the following two practical issues must be addressed:

- Claims or settlement conditions attested by a trusted third party: For example, as information of a traffic accident or personal safety may be recorded in the documents of a police agency, open data inquiry may be required via deregulation or rule adaption.
- Establishment of a gateway: A functional gateway, such as Oracle, is required to transfer external open data into the claims contract for on-chain use.

An alternative gateway design to achieve data exchange is feasible by using an observer pattern with the event-driven mechanism of smart contracts. The method can be configured as follows. The client of a claims contract registers the requested data category in the open data server. Automatic notifications to the contract client are executed when the requested data are available. The contract client may forward data to the claims contract. Similarly, the contract client may inversely transfer status changes via emitting events, which carry status values/parameters before or after an incurred event, to the client and then to external systems. Common tokens (ie, a kind of cryptocurrency, such as Ether in the Ethereum platform) could be used as payment tools among contract counterparties; however, this requires a sound exchange mechanism between flat currency and tokens. If tokens are sensitive to exchange rate fluctuations, it may be difficult to build up a stable and widely adopted industrial ecosystem. Thus, the issuance of a specific token for the home care industry may be required for function as a payment instrument, thereby solving the issue that existing cryptocurrency is currently unavailable for commercial transactions.

Blockchain and Smart Contract–Enabled Applications in the Home Care Industry

We integrated the blockchain-based home care system with the existing home care service-matching engine. The matching between caretakers and caregivers was completed by the existing matching engine while matching results were implemented and utilized by the proposed blockchain system. Focused issues and functionalities of the integrated system are addressed and illustrated in Figure 2. This study identified major issues in the existing system, including traceability, timeliness, interoperability, and cost, and later, it investigated how these issues would be mitigated by innovated/renovated blockchain-based home care services.

Traceability

A lack of an adequate workforce may result from labor conditions, social image, and individuals' expectations. Less than 30% of trained caregivers choose to stay in the home care industry, and the lack of an audit system with openness and trust is a major home care issue [32]. Consequently, the ineffective management of caregivers' promotion mechanism and care center classification results in the potential loss of the home care workforce and poor trustworthiness of care centers. Therefore, the emphasis should be on recording and maintaining related data for establishing a better auditing/evaluation system. These data should be open access and be characterized by trust and immutability. The blockchain-based system in this study may provide better system trustworthiness compared with that of the existing system.

Timeliness

Home care assignments are dispatched to caregivers by manual notifications while the care recipients are informed of assignments later. Assignment procedures involve multiple participants with delays in information transmission and disputes regarding assignment fairness. This study utilized Solidity, a programmable language provided by the blockchain platform, along with an event-driven mechanism to formulate an automatic notification function [16] to reduce user queries and pending durations induced by the existing process. Better timeliness was achieved on service assignment and case notifications in the new system. Moreover, at the beginning of the home care service, smart contracts may activate insurance application/cancellation to protect caregivers' rights. The timely notifications and insurance-related process automation allow more fluent operation and monitoring of service processes.

Interoperability

Care centers have long relied on siloed databases to manage service information. With individual matching methods, the home care industry lacks uniform data exchange formats/standards, thereby increasing switching costs [23]. Additionally, a lack of standard operating procedures has resulted in the use of various auditing/evaluation indices, which causes discrepancies in auditing procedures. Based on the above concerns, the proposed system presents a standard interface for data communication and allows matching results to be uploaded from the existing matching engine. From the proposed standard procedure for service assignment and care service, the system can collect equivalent information for assessment indices and allow the unification of the performance evaluation of care centers and caregivers.

Cost

Heavy reliance on manual operation of case assignment and notifications, and the lack of standard operating procedures may increase administrative and management costs. This study used a blockchain-based system and smart contracts to integrate the overall service processes. In doing so, all detailed tasks in processes were automatically completed by smart contracts and a great reduction in operational cost was achieved by virtue of

this distributed process automation scheme enabled by smart contracts.

Recap of the Advantages of Adopting Blockchain-Based Home Care Services

The results of comparative analyses between the blockchain-based system and the existing system clearly shed light on the general influences and potential benefits of a blockchain-based home care service system. Apart from offering better transparency and traceability that a blockchain system can achieve in common use cases, the distributed working paradigm facilitates data exchange among siloed databases of care providers, thus enhancing interoperability and reducing overall cost. Nevertheless, the findings from the questionnaire survey also provide nascent evidence on system feasibility and potentials from the individual stakeholder's perspective.

Conclusion

System Feasibility Based on Validation Results

The proposed system has the potential to enable task assignment and status tracking, decision control by the caretaker/caregiver, service notifications to relevant participants, and automatic insurance application/cancellation for caregivers. In the proposed system, the required core functions were implemented using four smart contracts, and they facilitate process automation by connecting related contracts. Additionally, a clear distinction between responsibility and accountability was achieved since decision control with the required information is granted to the caregiver/taker through the event-driven mechanism of smart contracts. According to the screenshots excerpted from the Remix implementation, validation indicators set prior to implementation were met exactly, thus proving system feasibility. Moreover, research findings from the questionnaire survey implied improved performance and excellence with the proposed system when compared with the existing system. Therefore, a blockchain-based home care service system may have potential for future applications.

Future Prospects

Based on the validation of system feasibility, future studies should pay more attention to the improvement of system effectiveness and operational management. Technically, blockchain, as a distributed ledger system, has a revocable feature once data are added to the chain. Therefore, data structures and computing algorithms for the implemented smart contracts are required for scalability. From a managerial perspective, an open and impartial auditing mechanism should be created under evaluation of the care center and caretakers. The evaluated information must be stored on the blockchain with further classification of caregiver and wage standards to formulate a positive cycle. Caregivers may therefore have more economic incentive to provide better home care service. This evaluation mechanism can be made more reliable by virtue of blockchain's open and tamper-proof characteristics. Evaluation information is open access to home care system participants such as caregivers, caretakers, and care centers. With better understanding of how service providers are assessed under specific standards, caregivers may have a better chance to accept or improve their services.

Acknowledgments

This research was supported by the Ministry of Science and Technology, Taiwan under contract number MOST-106-2221-E-005-053-MY3.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed methods of system design.
[DOCX File , 416 KB - medinform_v8i7e15472_app1.docx]

Multimedia Appendix 2
System implementation.
[DOCX File , 1299 KB - medinform_v8i7e15472_app2.docx]

Multimedia Appendix 3
Step-by-step workflow comparisons between the existing and proposed systems.
[PDF File (Adobe PDF File), 94 KB - medinform_v8i7e15472_app3.pdf]

Multimedia Appendix 4
Home care service system questionnaire.
[DOCX File , 29 KB - medinform_v8i7e15472_app4.docx]

References

1. Chang SE, Liu AY, Jang YT. Exploring employee's trust and organizational information monitoring for information security management. United States: IEEE; 2017 Presented at: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); October 14-16, 2017; Shanghai, China p. 1-3. [doi: 10.1109/cisp-bmei.2017.8302319]
2. Sharma U. Blockchain in healthcare: Patient benefits and more. IBM Blockchain Blog. 2017. URL: https://www.ibm.com/blogs/blockchain/2017/10/blockchain-in-healthcare-patient-benefits-and-more/ [accessed 2019-05-05] [WebCite Cache ID 78dvf0G02]
3. Nakamoto S. Bitcoin: A Peer-to-Peer Electronic Cash System. Bitcoin. 2008. URL: https://bitcoin.org/bitcoin.pdf [accessed 2019-05-05] [WebCite Cache ID 78dvf0G0K]
4. Buterin V. The meaning of decentralization. Medium. 2017. URL: https://medium.com/@VitalikButerin/the-meaning-of-decentralization-a0c92b76a274 [accessed 2019-05-05] [WebCite Cache ID 78dvf0Fzl]
5. Bhaskar ND, Chuen DL. Bitcoin mining technology. In: Chuen DL, editor. Handbook of digital currency: bitcoin, innovation, financial instruments, and big data. London: Academic Press; 2015:45-64.
6. Pongnumkul S, Siripanpornchana C, Thajchayapong S. Performance analysis of private blockchain platforms in varying workloads. United States: IEEE; 2017 Presented at: 26th International Conference on Computer Communication and Networks (ICCCN); July 31-August 3, 2017; Vancouver, Canada p. 1-6. [doi: 10.1109/icccn.2017.8038517]
7. Swan M. Blockchain: Blueprint for a New Economy. Sebastopol, CA: O'Reilly Media; 2015.
8. Park YR, Lee E, Na W, Park S, Lee Y, Lee JH. Is Blockchain Technology Suitable for Managing Personal Health Records? Mixed-Methods Study to Test Feasibility. J Med Internet Res 2019 Feb 08;21(2):e12533 [FREE Full text] [doi: 10.2196/12533] [Medline: 30735142]
9. Vazirani AA, O'Donoghue O, Brindley D, Meinert E. Implementing Blockchains for Efficient Health Care: Systematic Review. J Med Internet Res 2019 Feb 12;21(2):e12439 [FREE Full text] [doi: 10.2196/12439] [Medline: 30747714]
10. Azaria A, Ekblaw A, Vieira T, Lippman A. MedRec: Using Blockchain for Medical Data Access and Permission Management. United States: IEEE; 2016 Presented at: 2nd International Conference on Open and Big Data (OBD); August 22-24, 2016; Vienna, Austria p. 22-24. [doi: 10.1109/obd.2016.11]
11. Krawiec R, Filipova M, Quarre F, Barr D, Nesbitt A, Fedosova K, et al. Blockchain: Opportunities for health care. Deloitte. 2016. URL: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/public-sector/us-blockchain-opportunities-for-health-care.pdf [accessed 2019-05-05] [WebCite Cache ID 78dvf0G12]
12. Mettler M. Blockchain technology in healthcare: The revolution starts here. United States: IEEE; 2016 Presented at: IEEE International Conference on e-Health Networking, Applications and Services (Healthcom); September 14-16, 2016; Munich, Germany p. 14-16. [doi: 10.1109/healthcom.2016.7749510]
13. Szabo N. Smart contracts: Formalizing and securing relationships on public networks. First Monday 1997;2(9). [doi: 10.5210/fm.v2i9.548]
14. Franco P. Understanding Bitcoin: Cryptography, Engineering and Economics. New Jersey, NJ: John Wiley & Sons; 2014:68.
15. Buterin V. A next-generation smart contract and decentralized application platform. 2014. URL: https://ethereum.org/en/whitepaper/ [accessed 2019-05-05] [WebCite Cache ID 78dvf0Fzd]
16. Gavin W. Solidity. 2014. URL: https://solidity.readthedocs.io/en/develop/index.html [accessed 2019-05-05] [WebCite Cache ID 78dvf0Fzw]
17. Dikusar A. Smart contract benefits for business. XB Software Blog. 2017. URL: https://xbsoftware.com/blog/smart-contract-benefits-for-business/ [accessed 2019-05-05] [WebCite Cache ID 78dvf0Fzr]
18. Lewis A. Three common misconceptions about smart contracts. Bits on Blocks. 2017 Mar 7. URL: https://bitsonblocks.net/2017/03/07/three-common-misconceptions-about-smart-contracts/ [accessed 2019-05-05]
19. McFarlane C, Beer M, Brown J, Prendergast N. Patientory: A healthcare peer-to-peer EMR storage network. Patientory. 2017. URL: https://www.patientory.com/wp-content/uploads/2017/04/Patientory_Whitepaper-1.pdf [accessed 2019-05-05] [WebCite Cache ID 78fhMS7Vx]
20. Tian F. An agri-food supply chain traceability system for China based on RFID & blockchain technology. United States: IEEE; 2016 Presented at: 13th International Conference on Service Systems and Service Management (ICSSSM); June 24-26, 2016; Kung Ming, China. [doi: 10.1109/icsssm.2016.7538424]
21. Omaar J, Schwerin S, McMullen G. Forever isn't free: The cost of storage on a blockchain database. Medium. 2017. URL: https://medium.com/ipdb-blog/forever-isnt-free-the-cost-of-storage-on-a-blockchain-database-59003f63e01 [accessed 2019-05-05] [WebCite Cache ID 78dvf0G0Q]
22. Peterson K, Deeduvanu R, Kanjamala P, Boles K. A blockchain-based approach to health information exchange networks. United States: NIST; 2016 Presented at: NIST Workshop Blockchain Healthcare; September 26-27, 2016; Washington DC, United States p. 1-10.
23. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood) 2005;Suppl Web Exclusives:W5-10. [doi: 10.1377/hlthaff.w5.10] [Medline: 15659453]

24. Xu X, Pautasso C, Zhu L, Gramoli V, Ponomarev A, Tran AB, et al. The blockchain as a software connector. 2016 Presented at: Working IEEE/IFIP Conference on Software Architecture (WICSA); April 5-8, 2016; Venice, Italy p. 5-8. [doi: 10.1109/wicsa.2016.21]
25. Yue X, Wang H, Jin D, Li M, Jiang W. Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control. J Med Syst 2016 Oct;40(10):218. [doi: 10.1007/s10916-016-0574-6] [Medline: 27565509]
26. Gowda RG, Winslow LE. Collaboration between objects in OOA/OOD methodologies. 1993 Presented at: IEEE National Aerospace and Electronics Conference (NAECON); May 24-28, 1993; Dayton, USA p. 577-582. [doi: 10.1109/naecon.1993.290976]
27. Ratcliffe A. OOP needs OOA and OOD. Ratcliffe Technical Services. 2010. URL: http://www2.sas.com/proceedings/sugi25/25/ad/25p040.pdf [accessed 2019-05-25] [WebCite Cache ID 78dvf0G0c]
28. Schader M, Korthaus A, editors. The Unified Modeling Language: Technical Aspects and Applications. Berlin, German: Springer Science & Business Media; 2012.
29. Molecke R. How to learn Solidity: The ultimate Ethereum coding guide. Blockgeeks. 2017. URL: https://blockgeeks.com/guides/solidity/ [accessed 2019-05-05] [WebCite Cache ID 78dvf0G0E]
30. Chang SE, Chen YC, Wu TC. Exploring blockchain in international trade: Business process re-engineering for letter of credit.. Ind Manag Data Syst 2019 Aug 07:1712-1733. [doi: 10.1108/imds-12-2018-0568]
31. Chang SE, Chen YC, Lu MF. Supply chain re-engineering using blockchain technology: A case of smart contract based tracking process. Technol Forecast Soc Change 2019 Jul;144:1-11. [doi: 10.1016/j.techfore.2019.03.015]
32. Wang YT. Need and Development of Long-Term Care Workforce (in Chinese). CTCI Foundation. 2016. URL: https://reurl.cc/3Dq8a9 [accessed 2019-05-05]

Edited by G Eysenbach; submitted 12.07.19; peer-reviewed by J Lee, W Jiang, YR Park, JH Lee, JT te Gussinklo; comments to author 03.10.19; revised version received 26.11.19; accepted 03.06.20; published 28.07.20.

Please cite as:
Chang SE, Chen Y, Lu M, Luo HL
Development and Evaluation of a Smart Contract–Enabled Blockchain System for Home Care Service Innovation: Mixed Methods Study
JMIR Med Inform 2020;8(7):e15472
URL: https://medinform.jmir.org/2020/7/e15472
doi:10.2196/15472
PMID:32720903

Original Paper

Data Integration in the Brazilian Public Health System for Tuberculosis: Use of the Semantic Web to Establish Interoperability

Felipe Carvalho Pellison[1], MSc; Rui Pedro Charters Lopes Rijo[2,3,4,5], PhD; Vinicius Costa Lima[1], BSc; Nathalia Yukie Crepaldi[6], MSc; Filipe Andrade Bernardi[1], MSc; Rafael Mello Galliez[7], MSc, PhD; Afrânio Kritski[7], MSc, PhD; Kumar Abhishek[8], ME; Domingos Alves[5], MSc, PhD

[1]Bioengineering Postgraduate Program of the São Carlos School of Engineering, University of São Paulo, São Carlos, Brazil

[2]Polytechnic Institute of Leiria, Leiria, Portugal

[3]Institute for Systems and Computers Engineering at Coimbra, Coimbra, Portugal

[4]Center for Health Technology and Services Research, Porto, Portugal

[5]Department of Social Medicine of Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil

[6]Community Health Postgraduate Program, University of São Paulo, Ribeirão Preto, Brazil

[7]Academic Tuberculosis Program, Medical School of Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[8]Department of Computer Science and Engineering, National Institute of Technology, Patna, India

Corresponding Author:
Felipe Carvalho Pellison, MSc
Bioengineering Postgraduate Program of the São Carlos School of Engineering
University of São Paulo
Av. Trabalhador São-Carlense, 400
São Carlos, 13566-590
Brazil
Phone: 55 16 3373 9586
Email: felipecp@alumni.usp.br

Abstract

Background: Interoperability of health information systems is a challenge due to the heterogeneity of existing systems at both the technological and semantic levels of their data. The lack of existing data about interoperability disrupts intra-unit and inter-unit medical operations as well as creates challenges in conducting studies on existing data. The goal is to exchange data while providing the same meaning for data from different sources.

Objective: To find ways to solve this challenge, this research paper proposes an interoperability solution for the tuberculosis treatment and follow-up scenario in Brazil using Semantic Web technology supported by an ontology.

Methods: The entities of the ontology were allocated under the definitions of Basic Formal Ontology. Brazilian tuberculosis applications were tagged with entities from the resulting ontology.

Results: An interoperability layer was developed to retrieve data with the same meaning and in a structured way enabling semantic and functional interoperability.

Conclusions: Health professionals could use the data gathered from several data sources to enhance the effectiveness of their actions and decisions, as shown in a practical use case to integrate tuberculosis data in the State of São Paulo.

(JMIR Med Inform 2020;8(7):e17176) doi:10.2196/17176

KEYWORDS

health information systems; tuberculosis; ontology; interoperability; electronic health records; semantic web

Introduction

Background

One of the key issues regarding health information systems is their lack of interoperability [1]. Systems do not exchange data, and even when they do so, data do not have the same meaning. This situation can lead to disconnected service operations, rework, and poor comprehension of medical terms, which, in turn, can influence the quality of health services due to the increase in medical errors as well as health care costs [2]. In addition, data availability for scientific studies can be limited [3].

Tuberculosis (TB) is a curable disease, but in 2018, it was among the top 10 causes of death, with 1.5 million deaths and about 10 million new cases worldwide [4]. Brazil follows the directly observed treatment, short-course (DOTS) strategy recommended by the World Health Organization [5]. In Brazil, there are at least 8 main health information systems for TB, namely SISTB, ILTB, Hygia Web, Notification and Monitoring System for Cases of Tuberculosis in the State of São Paulo (TBWEB), Notification of Injury Information System (SINAN), Laboratory Environment Manager (sistema Gerenciador de Ambiente Laboratorial [GAL]), SITE-TB, and electronic Sistema Único de Saúde (e-SUS) AB. In some of these applications, health professionals have to reintroduce the same information and, when the patient's historic information is needed, manual processing is required. In a previous work, we studied the quality of the information among the data of 3 of these systems, and we found poor consistency and reliability [6].

In today's web, most of the available content is suitable for human interpretation and is therefore not easily accessible by other machines and systems. The Semantic Web, defined by Berners-Lee et al [7], can be specified as an extension of the current web, with the purpose of adding logic to the content to express the meaning of information, its properties, and the complex relationships existing between different types of data, so that it is possible to interpret the meaning of given data without worrying about the form of representation [8]. The goal is to create an efficient way to represent data on the World Wide Web to build a global database of connected data [9], through the semantic marking of web pages and existing relational databases using ontologies to provide a common meaning. Thus, this work proposes the use of the Semantic Web and a top-level ontology to support the interoperability of the Brazilian public health systems for tuberculosis, as an alternative to other standards, such as OpenEHR [10], mainly due to its flexibility, ease of implementation, and low level of needed intervention in the architecture of existing health information systems. Our work focuses on the specific scenario of Brazil for TB treatment and follow-up using DOTS.

Interoperability

The ability of two or more systems to exchange information and use it transparently is defined as interoperability [11]. For this, there are standards, languages, and protocols that must be followed, depending on the type of interoperability that one wishes to achieve.

According to the Healthcare Information and Management Systems Society, there are four levels of health information technology interoperability: foundational, structural, semantic, and organizational [12]. Foundational interoperability allows data exchange from one information technology system to be received by another and does not require the ability for the receiving information technology system to interpret the data. Foundational interoperability approaches range from direct connections to databases to service-oriented architectures using, for example, web services. Structural interoperability is an intermediate level that defines the structure or format of data exchange (ie, the message format standards) where there is uniform movement of health care data from one system to another such that the clinical or operational purpose and meaning of the data are preserved and unaltered. Structural interoperability defines the syntax of the data exchange. It ensures that data exchanges between information technology systems can be interpreted at the data field level. Structural interoperability is based on the concept of enterprise service buses, using standards for message formats. In health care systems, HL7 is the reference as a de facto standard [13]. Digital Imaging and Communications in Medicine (DICOM) [14] is another reference regarding the exchange of data between devices and image information systems. NextGen Connect Integration Engine [15] is also a cross-platform engine allowing the bidirectional sending of messages in many supported standards (such as HL7 V2, HL7 V3, HL7 Fast Healthcare Interoperability Resources, DICOM) between systems and applications. Semantic interoperability provides interoperability at the highest level, which is the ability of two or more systems or elements to exchange information and to use the information that has been exchanged. Semantic interoperability takes advantage of both the structuring of the data exchange and the data codification, including vocabulary, so that the receiving information technology systems can interpret the data. This level of interoperability supports the electronic exchange of patient summary information among caregivers and other authorized parties via potentially disparate electronic health record (EHR) systems and other systems to improve quality, safety, efficiency, and efficacy of health care delivery. OpenEHR [16] is a reference regarding semantic interoperability in health care. It is an open standard specification in health informatics that describes the management and storage, retrieval, and exchange of health data in EHRs. Finally, we can also consider organizational interoperability, which is concerned with how different organizations collaborate to achieve their mutually agreed electronic government goals. Concerned organizations need detailed agreement on collaboration and synchronization of their business processes to deliver integrated government services [17]. Contributing to the best practices of integrating systems and providing the basis for the organizational interoperability, Integrating the Healthcare Enterprise (IHE) profiles offer a common framework to understand and address clinical integration needs. IHE profiles are not just data standards; they describe workflows, which makes them more practical for use by health care information technology professionals and more applicable to their day-to-day activities [18].

Semantic interoperability can also be achieved with the Semantic Web. The basic blocks that define the Semantic Web are a standard data model, query protocol, and set of reference vocabularies. W3C standards and definitions, such as the Resource Description Framework (RDF), SPARQL Protocol, and RDF Query Language, and ontologies refer to these basic blocks, defined as a description language and data model, query protocol to obtain data stored in RDF, and formal representation of a given knowledge, respectively [19]. Ontologies can be defined as a formal representation of knowledge in a specific domain [20], aiming to formulate a rigorous and exhaustive conceptual scheme. In turn, Web Ontology Language is a semantic markup language for publishing and sharing ontologies, designed to describe classes and relationships between them [21].

In the Brazilian scenario, Ministry of Health Ordinance 2073/2011 regulates the use of interoperability standards in the scope of the Unified Health System (SUS) and the supplementary health sector [22] to guarantee functional and semantic interoperability for health information systems. Specifically regarding TB, follow-up of TB cases involves the filling of several registry instruments standardized by the Ministry of Health, such as the Individual Notification Form (Formulário de Notificação Individual [FNI]), Record of Directly Observed Treatment, and Record of Treatment and Follow-up of Cases of Tuberculosis, in addition to the electronic medical record, TBWEB, SINAN, SITE-TB, GAL, and e-SUS AB at a nationwide level. Other local systems are also involved, namely Hygia Web, SISTB [23], and hospitals' information systems, which are, respectively, governmental, state, and regional health information systems.

SISTB stores and centralizes information about the patient, treatments, examinations, and hospitalizations in the municipality of Ribeirão Preto. HygiaWeb is the public health management software of Ribeirão Preto city, which connects many levels of the local health care system. TBWEB is software developed by the government of the State of São Paulo for epidemiological surveillance. SINAN is software used nationwide to notify of every new case of certain compulsory notification diseases that are stipulated by the national government. GAL allows the management of routines and the monitoring of the steps for conducting exams, containing data that could be associated with the records of exams stored in the SISTB. SITE-TB is a platform that supports the notification of all prescriptions for treatment that does not involve the drugs that are generally used for drug-resistant TB (ie, rifampicin, isoniazid, pyrazinamide, and ethambutol). E-SUS AB is the basic attention health management software at a national scope.

These numerous systems with different technologies, data formats, and semantics generate difficulties in the follow-up of the patient since they can create duplicity, losses, and contradictory information [6].

In the next section, we present some key works in the area of interoperability using the Semantic Web.

Related Work

At first, Lopes and Oliveira [24] used the metaphor of closed and distributed silos where the health data are fragmented distributed, thus highlighting the rusticity of the software that deals with them. The authors suggested a migration to the Semantic Web model through a framework idealized by the same authors with several resources that allow not only the extraction of data but also knowledge. Valle et al [25], in a similar initiative, advocated for the adoption of the Semantic Web paradigm with technologies widely used in the market, such as Extensible Markup Language and web services, to achieve interoperability in health. The authors defended this approach as it promotes the combination of syntactic and semantic interoperability between applications.

Abhishek and Singh [26] proposed an ontology following the principles of the Basic Formal Ontology (BFO) for the Indian scenario of TB. Hitzler and Janowicz [27] emphasized the increasing use of the Semantic Web. Such growth is possible thanks to the application of the Semantic Web paradigm not being tied to a specific type of knowledge or area. On the contrary, the Semantic Web comes to support environments where interdisciplinarity and heterogeneity are implicit in the routine. In this sense, the increase in the number of conventions and people interested in the subject is justified, whether they come from the academic or commercial milieu.

Ogundele et al [28] developed an ontology for representing, consolidating, and structuring knowledge about the factors that influence treatment adherence behavior in patients with TB. The repository created can be used to find potential factors affecting TB treatment adherence in similar communities that were used in the study, generating its risk indices and also helping the monitoring of patients and their follow-up.

Heterogeneity between conceptualizations must be resolved before handling the term-level syntactic heterogeneity so that semantic interoperability can be conducted in an effective way. In the proof of concept by Gonçalves et al [29], an experiment was conducted that provided evidence that the electrocardiogram ontology can be effectively used to support the design of other interoperable versions that refer to electrocardiograms, such as the HL7 Annotated Electrocardiogram. The authors also affirmed, through their results, that their method can also be applied in other domains.

Kumar [30] discussed some vocabulary as well as how it can be used to achieve interoperability between applications. However, he also discussed the issues surrounding the privacy and interoperability of applications that use the Semantic Web. Also, according to Zenuni et al [31], mapping proprietary formats in ontologies is a complex and intense task, and the maintenance of ontologies is one of the delicate points.

Despite these challenges, some key practical case studies must be referenced. Belleau et al [32] presented the Bio2RDF project using Semantic Web tools to cluster biomedical knowledge extracted from numerous databases. McMurray et al [33] developed a conceptual model of regional clinical electronic exchanges between health care. The ontology allowed visualization of the model and instances by the computational

model and was used to validate a subset of the collected data using a different database, although it still had a related database with the interests of the research. Jiang et al [34] described the development of a tool to solidify the definitions recommended by the International Classification of Diseases, version 11. The classification generated by their work through a governance model that involved expert consensus, collaborative and distributed validation, and support allowed these parameters to be optimally tuned. These classifications were then compared with other values found in the Unified Medical Language System and the Systematized Nomenclature of Medicine - Clinical Terms. The compilation of the results was submitted to specialists for evaluation and effective assignment of the degree of usability. Finally, Abhishek and Singh [26] created an ontology to assist the decision-making of managers of the National Indian Tuberculosis Control and Management Program. The basis of this ontology was the BFO, characterized as a meta-ontology that allows the hierarchy and correct division and classification of the entities of the domain that one wishes to represent. This approach facilitates possible mapping between other ontologies that make use of the same meta-ontology for their construction as well as guarantees semantic consistency for application of knowledge extraction algorithms based on Semantic Web Rule Language. Several examples of queries were demonstrated in said work, proving the robustness of the solution adopted.

The next section presents the proposal of an interoperability solution for the Brazilian public health system supporting TB.

Methods

The cornerstone of a Semantic Web solution is its underlying ontology. Thus, the first step was the development of an ontology considering the clinical TB concepts; the different existent systems FNI, TBWEB, SINAN, SISTB, HygiaWeb, e-SUS AB, and GAL; and the concepts related with TB and DOTS. DOTS is the international strategy for TB control recommended by the World Health Organization that has been recognized as a highly efficient and cost-effective strategy. As already mentioned, this is the strategy adopted by the Brazilian government.

The entities of the ontology were allocated under the definitions of the BFO, which is a top-level ontology initially developed for use in scientific domains such as biomedicine. BFO sees reality regarding a top-level division of all particular entities (individuals) into the two disjoint categories of continuant and occurrent. Continuant entities include objects, attributes, and locations and are contrasted with occurrence entities, which include processes and temporal regions. Processes happen in time and so have temporal parts. Continuants, in contrast, exist in full at any time in which they exist at all. Because it is an upper-level ontology of a fundamentally realistic methodology and has a high level of representation, it allows the mapping of several entities, processes, and their respective functions and characteristics within a temporal space, standing out over other ontologies that only take snapshots of these situations. Given our interest in mapping terms from both medicine and administrative areas, such top-level ontology is an excellent alternative. Figure 1 presents the resultant ontology for DOTS in the Brazilian health policies scope, and Figure 2 represents its object properties.

Figure 1. The ontology for directly observed treatment, short-course (DOTS) to support interoperability in the Brazilian public health system.

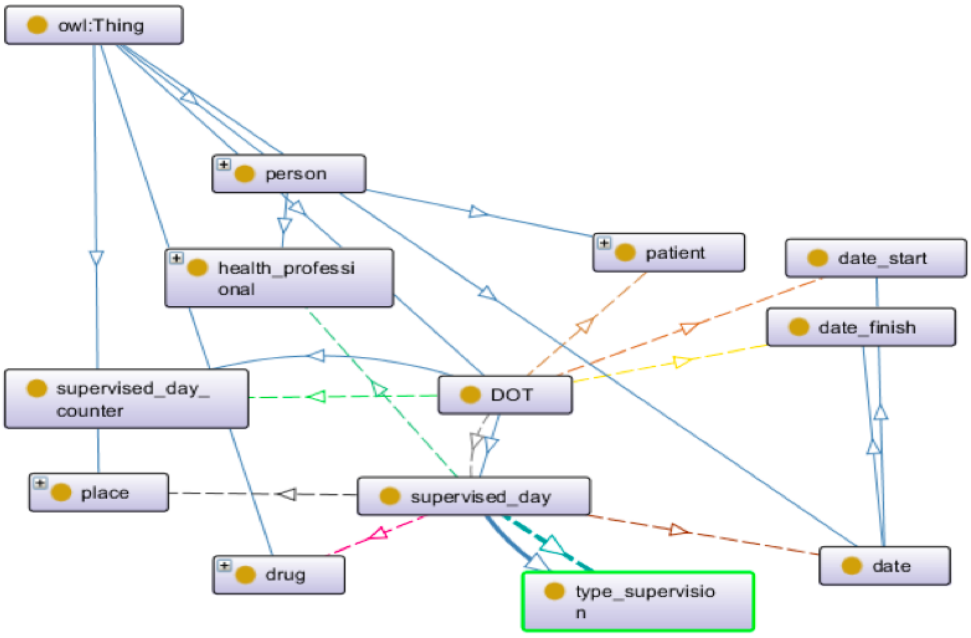
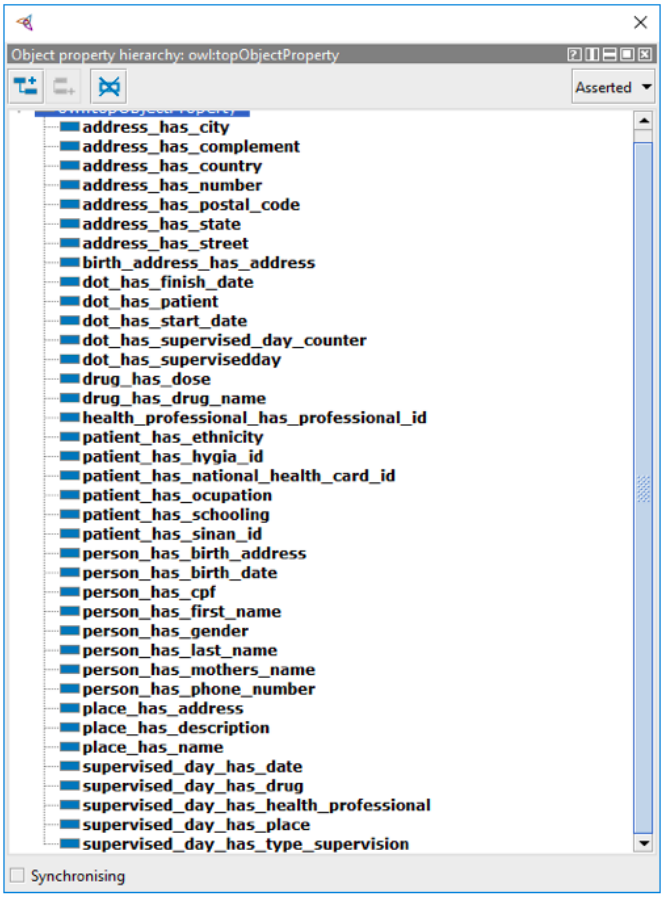


Figure 2. Object properties related to directly observed treatment, short-course (DOTS).



Abhishek and Singh [26] demonstrated the creation of an ontology for the Indian scenario of TB, serving as the basis for the development of the ontology presented in this work. When using the same framework to represent the Brazilian scenario, the mapping ontology-ontology was facilitated. This is because BFO already classifies these terms, making the stage of meaning abstraction trivial and necessitating only the relation of the terms with similar meanings. By eliminating the step of meaning abstraction, an inherent subjectivity burden that can lead to errors in the mapping of one ontology to another, is also removed. Figures 3-5 are excerpts of the concepts that were mapped into BFO.

The mapping of terms in the BFO structure presented considerable difficulty, given the philosophical complexity involved in the definitions and the degree of comprehensiveness we chose to take. It is clear that the granularity of the actions specified in the construction of the ontology can grow as the engineer wants. However, for this work, the resulting formalization shown in Figures 3-5 was the result of consensus among the authors and judged sufficient to support interoperability between the proposed systems.

Figure 3. First part of the tuberculosis entities mapped into the Basic Formal Ontology (BFO).

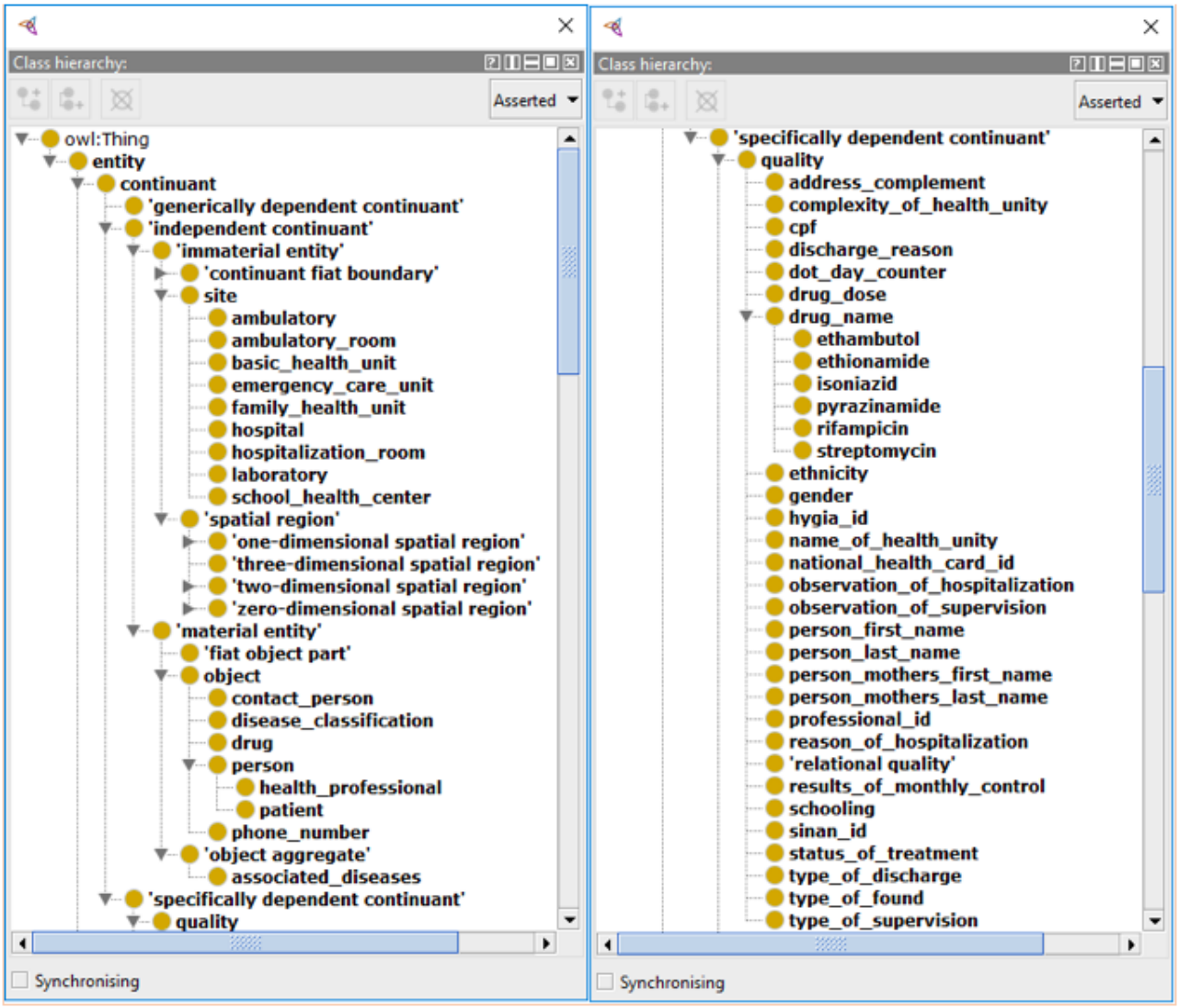


Figure 4. Second part of the tuberculosis entities mapped into the Basic Formal Ontology (BFO).

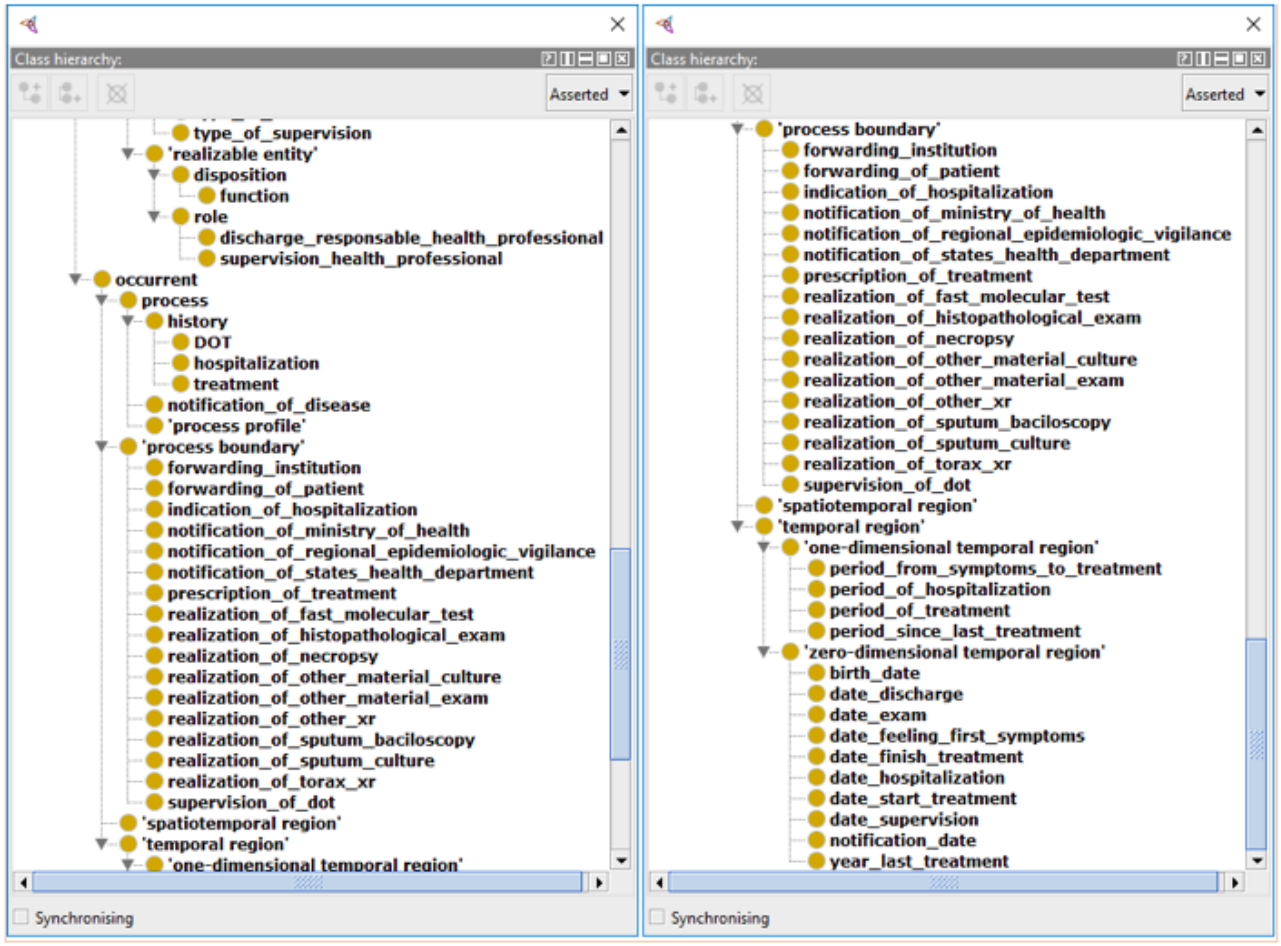
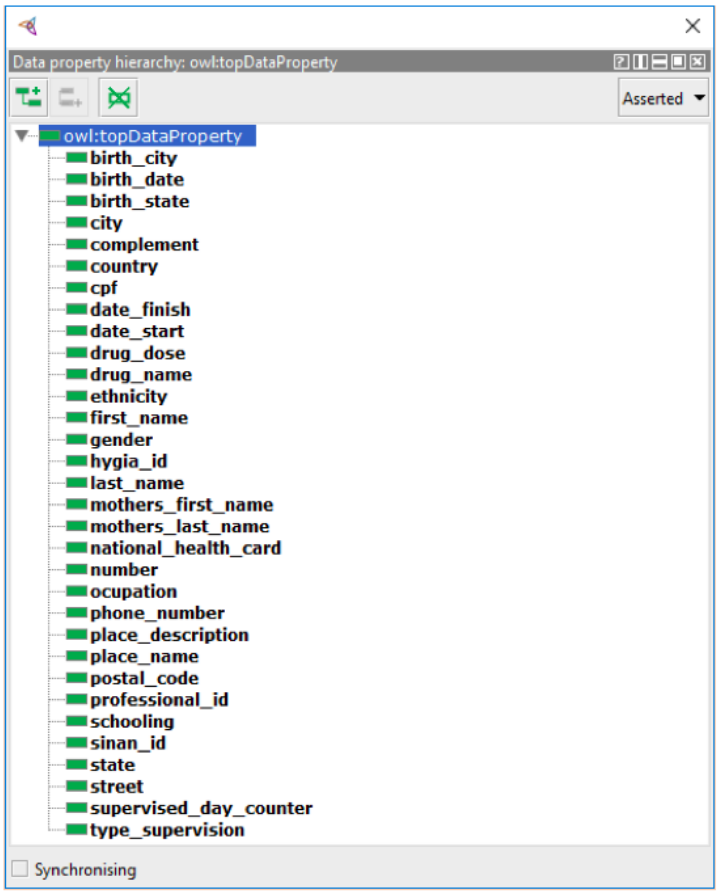


Figure 5. Translated terms that allowed the interchange of data between the applications.



The key concepts to achieve interoperability between systems and the recovery of DOTS data are described in Figure 5. Patients' personal data (address, birth date, Natural Persons Register [Cadastro de Pessoas Físicas, CPF], mother's name) and health-related information (national health card, HYGIA ID, SINAN ID) were described. The transitivity characteristics of the retrieved data, together with the semantic interoperability guarantee provided by the ontology, allowed the merging of the patients' data from the participating applications. This increases the relevance of the information available to the health professional who needs to make strategic and clinical decisions.

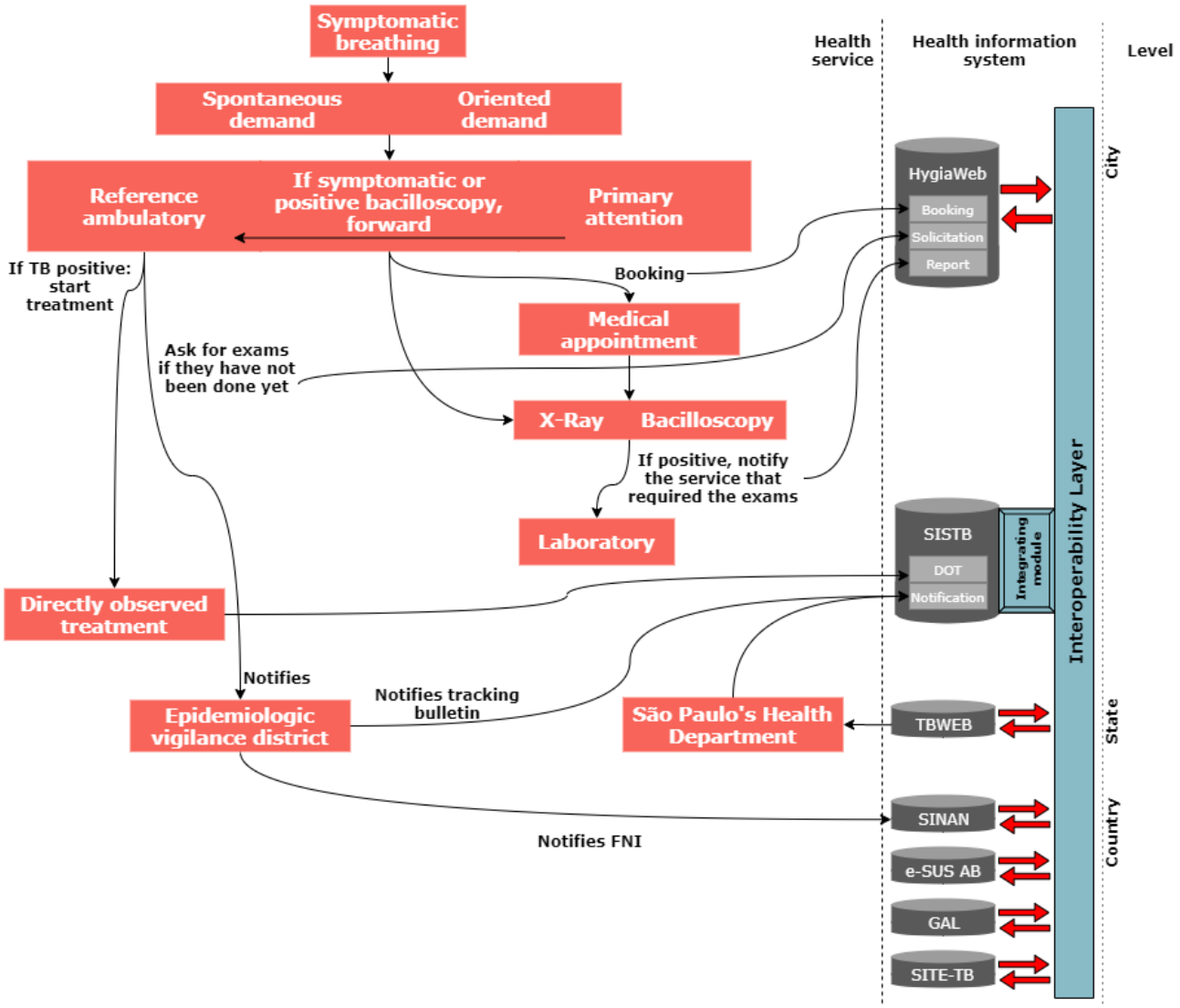
This ontology allows the marking of all the necessary data to reach interoperability between the systems previously mentioned.

Data can be obtained through queries on endpoints that support the SPARQL language, as well as any data source semantically marked with the ontology (websites, text documents, spreadsheets, RDF). In the latter case, however, there is a need to process these data to fully exploit the information that does not occur in data returned directly by the SPARQL endpoints. A challenging issue concerns the compatibility of this proposed solution with legacy databases. Such concern is justified by the immeasurable value of knowledge accumulated during several years of care in health services. In this sense, an alternative was sought to address performance and compatibility concerns with relational database management systems. The tool chosen was the D2R Server [35,36], which allows the establishment of a virtual database based on a given ontology and the execution of SPARQL queries on the legacy database, returning the desired information, which are the final values itself, in RDF format and used in a Semantic Web paradigm. This solution, invoked via an application programming interface, contributes in a positive way to reduce the impact caused by the paradigm shift from legacy systems to the Semantic Web. With this approach, there is no need to treat the whole database so that it becomes usable and consumable by web-based semantic applications. Then, this framework allows data exchange between legacy and Semantic Web–enabled applications.

Results

In Figure 6, the implementation of the interoperability layer between FNI, TBWEB, SINAN, SISTB, HygiaWeb, e-SUS AB, SITE-TB, and GAL, which allows transparent information exchange is presented. The interoperability layer was based on the Semantic Web paradigm and standards preconized by W3C. This paradigm allows the extraction of content from these systems in an optimized way for machines through a web service, opening a range of possibilities for generating useful information for decision making, as specified by Berners-Lee et al [7].

Figure 6. Information flow for tuberculosis treatment after implementation; adapted and improved from [37,38]. DOT: directly observed treatment; e-SUS: electronic Sistema Único de Saúde; FNI: Formulário de Notificação Individual; GAL: sistema Gerenciador de Ambiente Laboratorial; SINAN: Notification of Injury Information System; TBWEB: Notification and Monitoring System for Cases of Tuberculosis in the State of São Paulo.



The search for health services occurs when the patient has some symptoms related to the respiratory system. Such demand may be spontaneous or oriented. Spontaneous demand refers to when the patient searches for a general practitioner or pneumologist on his or her own. A demand-oriented search is when the patient is referred to this specialty by the general practitioner, primary care physician, or family physician. The patient is then referred to primary care (if the search for care has not started at this point). If the symptoms corroborate the presumed TB, chest X-ray and sputum examination are requested, and a medical appointment is scheduled. After obtaining the exam results, the patient is referred to an outpatient clinic where the treatment will begin. Treatment is directly observed, and all follow-up data are stored in SISTB. The directly observed treatment is a health policy that intends to closely monitor the evolution of the treatment and patient to increase the effectiveness and success rate of the treatment.

To achieve interoperability, it is necessary to semantically tag participant systems using the ontology previously presented. To do that, HTML pages can be tagged via a Microdata framework [39,40] that extend the HTML specification with specific attributes. Also, a middleware, such as D2R Server, can be configured to expose relational databases as virtual RDF datasets.

Tags refer directly to the terms used in the ontology. This markup format was chosen because it allows the search engines to easily extract knowledge from the fields marked with the tags since the HTML language is a common basis for web applications. The great advantage of this knowledge extraction is that semantic interoperability is already implicitly inserted in the page context page, since the ontology gives every logical structure to which the tagged data is linked, avoiding further work of assigning meaning to the data returned.

In each of those systems, an active SPARQL endpoint service is desirable to allow running SPARQL queries on information stored in legacy databases. Such an endpoint can be provided by middleware, like the D2R Server. This is fundamental for extracting data that have been stored before adopting the Semantic Web paradigm to enable the interoperability layer.

Extracting information directly from marked pages is done through the library Any23, which directly extracts the objects

or literals and their tags (corresponding to the ontology). With the extracted data, it is possible to realize several types of SPARQL queries and to incorporate such information in its local database for any queries.

A very simple SPARQL query, where all properties of all patients are returned, is being used as the basis for the incorporation of data extracted from other systems marked with the respective ontology.

It is important to note that this query can be executed on the data of all the marked systems. This guarantees that the returned data have the same meaning, since they were marked with the same ontology and are, therefore, interoperable between the systems.

Such an approach allows data tagged with this same ontology to be fetched on any system (via HTTP requests or SPARQL endpoint). The returned data are then treated and incorporated into the system, if desired, or can be used for ad hoc queries and statistics, allowing rapid decision making by health professionals. In this sense, both semantic interoperability and functional interoperability work in function of data integrality and rapid response sought by health information systems.

All health information systems in the scope of this work have common data identifying the patient (CPF, National Health Card, birth date). That is, with the semantic marking, it is possible to return the data referring to a particular patient and to aggregate them in a single result or to import only enough data for the decision making, as shown in Figure 7. The intersection of these records represents a snapshot of the current patient's health situation, corroborating with the integrality in the patient's health care. The retrieval of this information from several health information systems can allow health professionals to make their decisions with more details than would be present if they were using an isolated system. From the managerial point of view, data aggregation for the accomplishment of demographic studies is also improved. The development of more specialized public health policies with better effectiveness is also facilitated since the availability of information on most patients is increased.

Figure 7. Results of a SPARQL query simultaneously on SISTB and Notification and Monitoring System for Cases of Tuberculosis in the State of São Paulo (TBWEB) for a specific patient with sinan_id=10.

Results of SPARQL Query over SisTB

propriedade	valor
<http://sistb-dev.ddns.net/ontology/Paciente/nroProntuario>	"6629270"
<http://sistb-dev.ddns.net/ontology/Paciente/nroSinan>	"10"
<http://sistb-dev.ddns.net/ontology/Paciente/idade>	"17"
<http://sistb-dev.ddns.net/ontology/Paciente/dataNascimento>	"30/04/2000"
<http://sistb-dev.ddns.net/ontology/Paciente/gestante>	"Não"
<http://sistb-dev.ddns.net/ontology/Paciente/nomePaciente>	"NOME TESTE 10"
<http://sistb-dev.ddns.net/ontology/Paciente/cartaoSus>	"80000105"
<http://sistb-dev.ddns.net/ontology/Paciente/nomeMae>	"MÃE TESTE 10"
<http://sistb-dev.ddns.net/ontology/Paciente/genero>	"Masculino"
<http://sistb-dev.ddns.net/ontology/Paciente/escolaridade>	"0 ano(s) concluido(s)"
<http://sistb-dev.ddns.net/ontology/Paciente/etnia>	"Pardo"
<http://schema.org/Person/telephone>	"(16) 1051-0510"
<http://schema.org/Person/address>	"Endereço: Rua 105Cidade: RIBEIRAO PRETOUF: SP"

Results of SPARQL Query over TBWeb

propriedade	valor
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"RX Outro"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Normal"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"Cultura do Escarro"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Em andamento"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"Baciloscopia de escarro"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Em andamento"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Amicacina (A)"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"Outros"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"dd"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Ofloxacina (Ofl)"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Outras Drogas"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"Necropsia"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Não sugestivo TB"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Pirazinamida (Z)"
<http://sistb-dev.ddns.net/ontology/Medicamento/tomou>	"Sim"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Etambutol (E)"
<http://sistb-dev.ddns.net/ontology/Medicamento/tomou>	"Sim"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	Isoniazida (H)"
<http://sistb-dev.ddns.net/ontology/Medicamento/tomou>	"Sim"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Clofazimina (Clo)"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"Baciloscopia Outro Material"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Em andamento"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"RX de Tórax"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Normal"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"Cultura do Escarro"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Positivo"
<http://sistb-dev.ddns.net/ontology/Paciente/nroSinan>	"10"
<http://sistb-dev.ddns.net/ontology/Paciente/genero>	"Masculino"
<http://sistb-dev.ddns.net/ontology/Paciente/dataNascimento>	"30/04/2000"
<http://sistb-dev.ddns.net/ontology/Paciente/nomePaciente>	"NOME TESTE 10"
<http://sistb-dev.ddns.net/ontology/Paciente/ocupacao>	"Estudante"
<http://sistb-dev.ddns.net/ontology/Paciente/ocupacao>	"Outra"
<http://sistb-dev.ddns.net/ontology/Paciente/nomeMae>	"MÃE TESTE 10"
<http://sistb-dev.ddns.net/ontology/Paciente/escolaridade>	"De 8 a 11 anos"
<http://sistb-dev.ddns.net/ontology/Paciente/idade>	"17"
<http://sistb-dev.ddns.net/ontology/Paciente/cartaoSus>	"10"
<http://sistb-dev.ddns.net/ontology/Paciente/cpf>	"10"
<http://sistb-dev.ddns.net/ontology/Paciente/rg>	"10"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Terizidona (T)"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Estreptomicina (S)"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Etionamida (Et)"
<http://purl.org/dc/terms/title>	"TBWEB"
<http://sistb-dev.ddns.net/ontology/Medicamento/nome>	"Rimfampicina (R)"
<http://sistb-dev.ddns.net/ontology/Medicamento/tomou>	"Sim"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"Anti-HIV"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Em andamento"
<http://sistb-dev.ddns.net/ontology/Exame/nome>	"Histopatologico"
<http://sistb-dev.ddns.net/ontology/Exame/status>	"Não sugestivo TB"

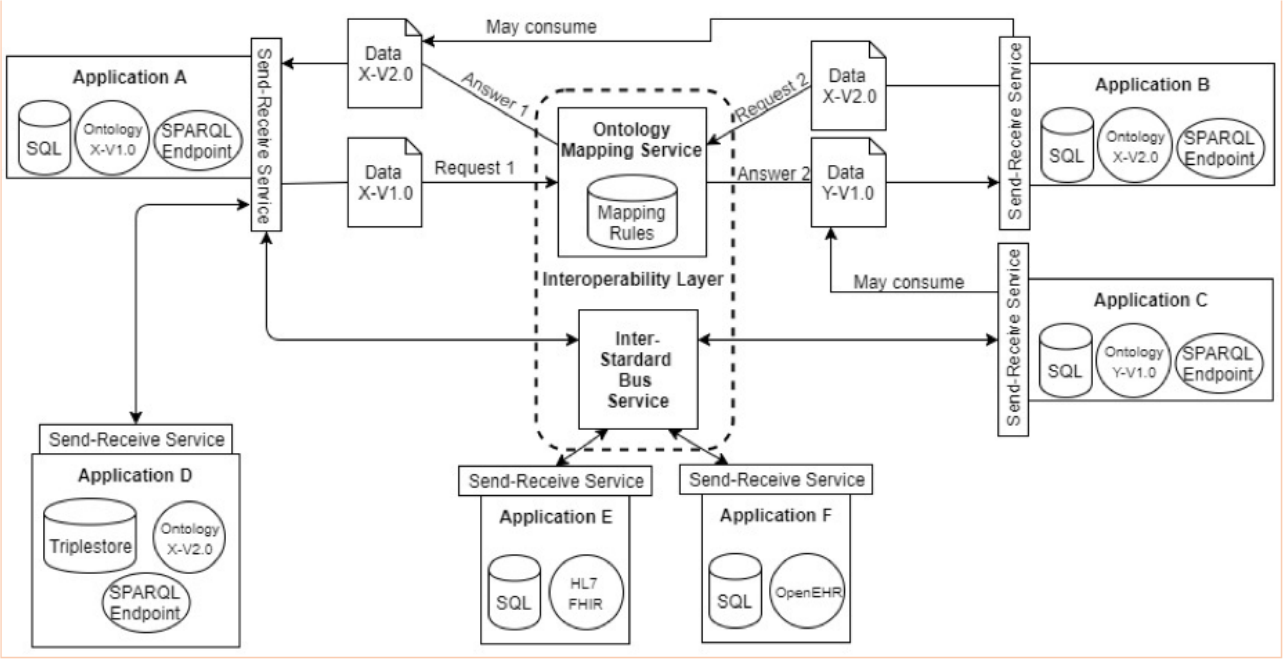
A key challenge is to reach interoperability with other standards, such as HL7, OpenEHR, and IHE profiles. In the case of HL7, a viable approach is deploying a middleware capable of translating extracted data through semantic markup into HL7 messages (V2.x or V3.x). Plastiras and O'Sullivan [41] published their work with similar development of a middleware to interoperate data from personal health records to EHRs. However, our proposed middleware receives as input the extracted data of semantically tagged applications, constructs the messages, and forwards them to the recipients. In this scenario, an effort must be made to map the entities of the domain-specific ontologies in the pre-specified fields recommended by the HL7 standard.

Such an effort would be similar to the mapping of which entities match a given previously specified OpenEHR archetype. That is, to interoperate systems that use the Semantic Web with systems that use the OpenEHR standard, it is necessary to ensure that the archetypes are fully represented by entities of an ontology. The reciprocal is also true since the mapping of

archetypes into entities of an ontology is also extremely necessary. Such processes are exemplified in Figure 8.

As a recent result of the following proposed architecture, it is valid to cite the work of Pellison et al [42]. Their work presented a proof of concept regarding TB data integration in the State of São Paulo, Brazil, using Semantic Web resources, such as SPARQL queries and RDF. Throughout a federated query, data were simultaneously obtained from TBWeb, the state governmental system, and SISTB, the regional TB information system used mainly in the city of Ribeirão Preto (Brazil). By doing this, it was possible to combine data from both sources with aggregated semantic value.

Figure 8. Architecture for interoperability between applications that use the Semantic Web and other standards (ie, HL7 FHIR and OpenEHR).



Demographic data were used to push points on a map and to compare values among datasets, which included latitude, longitude, pregnancy situation, age, gender, city of notification, federative unity (state), schooling, and ethnicity. Users were able to obtain information about notification of TB cases using filters available in the interface of the map. The search result was drawn as a heatmap, according to the municipalities that have notified TB cases.

This work, as a proof of concept, calls attention to the importance of working in solutions that could improve the quality of data in the health field and daily activities of health professionals.

Discussion

Principal Findings

This study presents an ontology based on the BFO meta-ontology to support TB-related data in Brazil. The construction of the BFO-based ontology benefits from the fact that it is easily related to other ontologies built on the same framework. This allows different ontologies to be related in an easier way because their entities are organized systematically and hierarchically according to their semantic meaning advocated by the same meta-ontology. This interontology relation has great potential because it allows marked data to be shared among the institutions by carrying their respective semantic values and broadening their potential of multicentric research. Currently, there are initiatives to map terms and construct ontologies related to the treatment of TB in other scenarios, as quoted by Abhishek and Singh [26]. As the present work was carried out based on the same meta-ontology and due to the interontology relationship, the possibility of data transposition between the marked systems is feasible and can be interesting for both parties, thus allowing semantic interoperability.

Although solutions like OpenEHR makes domain semantics a central concern, it is optimized to provide a data platform with a stronger focus on the persistence of data, with data exchange a secondary focus. OpenEHR uses a large set of complex archetypes (ie, the model or pattern) for the capture of clinical information, which are designed to provide a maximal set of data elements. This breadth and depth inevitably bring a level of complexity. Semantic Web, on the other hand, delivers more flexibility through the use of ontologies as models to represent health domain data. As OpenEHR archetypes, ontologies can be reused, extended, and adapted to specific demands (*mutatis mutandis*) [43] and can easily be applied to a health information system without so much effort. By using tools designed for the Semantic Web to add meaning to data, such as virtual graph (or tuples) repositories, and developing integrated application programming interfaces to perform data exchange, the level of intervention in health information systems is reduced, resulting in more immediate benefits for functional and semantic interoperability.

During the development of the architecture, some challenges and possibilities of fruitful future research were found. All ontologies can be modified, either because of changes in the domain definition or because corrections in their construction

and other adjustments considered necessary for the proper functioning of the ontology. However, such changes directly impact the systems marked with it, and special attention is needed for the data that are marked by the ontology. It is necessary to control ontology versions to admit reconstruction of the ontology cycle of life and to trace the meaning of the extracted data in the version in which it was marked. Therefore, it is necessary to find solutions for the readjustment, most easily and intuitively as possible, for data re-marking and to minimize negative impact on the systems. Some work has already been developed in this area, focusing on the prediction of patterns of ontology changes, for instance, as demonstrated by Javed et al [44].

With version control, we also need to pay attention to how to mark these data, since this task requires considerable effort in systems that use ontologies with many entities and terms. Such automatic markup has a small amount of work performed and documented, perhaps justified by the need for specialized markers in the domain for which the respective ontologies were developed. It is still possible to emphasize some initiatives that have been proposed to help developers to carry out the marking in, at least, a semi-automatic way. Among them, we can mention UCCA-App [45], MnM [46], and SemTag and Seeker [47]. The reduction of the markup effort, be it at first or after the publication of a new version of a given ontology, should be a focus of future work, to provide simplified system maintenance that uses the Semantic Web paradigm.

The flexibility in constructing ontologies and the high level of abstraction they possess makes the mapping process for archetypes relatively trivial when one already has an accepted archetype for the foundation responsible for managing the standard repository. It is still important to emphasize that the Semantic Web approach can be incorporated into several IHE profiles so that integration can occur in several areas, being enough to have an ontology that supports all of them and their respective processes.

Interoperability protocol initiatives that are being carried out in Brazil, especially by the Ministry of Health, have a lot of bureaucracy and technical challenges. This means that the implementation of these standards by themselves includes a lot of effort in recoding the already existing applications, the deployment of endpoints to allow functional interoperability, and having the mapping of the scenario accepted by the institution that takes care of the standard (ie, OpenEHR). This means that the time that is necessary to make many health applications interoperable, both semantically and functionally, is much longer than in the Semantic Web approach. This affirmation is supported by the concept of the Semantic Web itself, which allows a much more flexible governance to develop its ontologies. Semantic Web usage has the advantage of having a much more dynamic domain scope governance than other interoperability standards such as the previously cited HL7 and OpenEHR. Such dynamics favor the evolution of terms and adaptation to new trends, something that is inherently recurrent in health, where techniques, new procedures, and clinical protocols evolve every day.

With this architecture in mind, it is possible to expand the initiative of Conecta SUS, established by the Strategic Plan [48] made by the National Health Terminology Center from the Brazilian government. Conecta SUS is an alternative to improve the scenario shown by Rijo et al [49], where the authors show and assess the lack of interoperability through many health institutions in Brazil.

The results present a viable and practical use of this architecture, opening a new horizon of application at any health care level or specialty.

Conclusions

In this work, research and implementation of an ontology that supports the Brazilian scenario of TB were conducted. Such an ontology was constructed using the meta-ontology BFO for the classification of terms. This formalization will allow the mapping of Brazilian ontology entities for TB among other ontologies that also have used BFO as a model and have similar entities, facilitating semantic interoperability. Upon the ontology that was built, an architecture was developed to allow functional interoperability between applications that store health data related to TB. Applications were marked via microdata attributes with the terms of the ontology created. This markup enables content extraction from multiple applications from a single SPARQL query on the endpoints installed in each application. It is worth mentioning that solutions have been implemented to run SPARQL queries in relational databases and triplestores, thus allowing the maintenance of legacy databases. The example presented in this article shows how data from a given patient can be returned from all applications that had the same enrollment. The returned data maintained their meanings, and semantic and functional interoperability was achieved. A limitation of our work is the mapping of the entities that concern multidrug-resistant TB, extensively multidrug-resistant TB, and the notification application for these cases (SITE-TB) and comorbidities. Future work will include the map of these workflows and other support applications like the National Regulation System (SISREG) and demographic applications from the Brazilian Institute of Geography and Statistics (IBGE).

Despite the ease of using legacy databases, there is a need to improve the services that would facilitate the implementation of this solution in daily practice. Automatic data marking can be an area of study interest, aiding in the effort required to attribute semantic meaning to the data. Other viable examples in the short term would be the implementation of an interontology and intra-ontology mapping service and also an interstandard message bus and routing service. The first one would allow data to be marked with more than one ontology version and consequently be consumed by different applications that use different ontologies. The second one would allow interoperability between applications that use paradigms and patterns other than the semantic web, such as HL7 FHIR and OpenEHR.

Acknowledgments

This research is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) of the Ministry of Science, Technology, Innovations and Communications (process: 440758/2018-1) and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), a foundation of the Ministry of Education of Brazil and the Ministry of Health (process: 88887.141211/2017-00).

Conflicts of Interest

None declared.

References

1. Gambo I, Oluwagbemi O, Achimugu P. Lack of Interoperable Health Information Systems in Developing Countries: An Impact Analysis. Journal of Health Informatics in Developing Countries. URL: http://www.jhidc.org/index.php/jhidc/article/view/60 [accessed 2020-02-03]
2. Iroju O, Soriyan A, Gambo I, Olaleke J. Interoperability in Healthcare: Benefits, Challenges and Resolutions. International Journal of Innovation and Applied Studies 2013;3(1):262-270 [FREE Full text]
3. Douglass K, Allard S, Tenopir C, Wu L, Frame M. Managing scientific data as public assets: Data sharing practices and policies among full-time government employees. J Assn Inf Sci Tec 2013 Nov 26;65(2):251-262. [doi: 10.1002/asi.22988]
4. Global Tuberculosis Report 2019. World Health Organization. URL: https://www.who.int/tb/global-report-2019 [accessed 2020-02-03]
5. Creswell J, Sahu S, Sachdeva KS, Ditiu L, Barreira D, Mariandyshev A, et al. Tuberculosis in BRICS: challenges and opportunities for leadership within the post-2015 agenda. Bull. World Health Organ 2014 Jun 01;92(6):459-460. [doi: 10.2471/blt.13.133116]
6. Yamaguti VH, Vicentine FB, de Lima IB, Zago L, Rodrigues LML, Alves D, et al. Data quality in tuberculosis: the case study of two ambulatories in the state of São Paulo, Brazil. Procedia Computer Science 2017;121:897-903. [doi: 10.1016/j.procs.2017.11.116]
7. Berners-Lee T, Hendler J, Lassila O. Book. The Semantic Web: a New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Scientific American; 2002:a.
8. Robu I, Robu V, Thirion B. An introduction to the Semantic Web for health sciences librarians. J Med Libr Assoc. ? 2006;94(2):205.
9. Laufer C. Guia de web semântica. Guia de web semântica. URL: https://nic.br/media/docs/publicacoes/13/Guia_Web_Semantica.pdf [accessed 2020-02-03]
10. Pahl C, Zare M, Nilashi M, de Faria Borges MA, Weingaertner D, Detschew V, et al. Role of OpenEHR as an open source solution for the regional modelling of patient data in obstetrics. Journal of Biomedical Informatics 2015 Jun;55:174-187. [doi: 10.1016/j.jbi.2015.04.004]
11. Institute of Electrical and Electronics Engineers. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. Website 1991 Jan 18:A. [doi: 10.1109/IEEESTD.1991.106963]
12. HIMSS. What is Interoperability? Healthcare Information and Management Systems Society. Published December 9. 2019. URL: https://www.himss.org/what-interoperability [accessed 2020-02-03]
13. About Health Level Seven International | HL7 International. Health Level Seven International. URL: http://www.hl7.org/about/index.cfm?ref=nav [accessed 2020-02-03]
14. DICOM. DICOM Standard. URL: https://www.dicomstandard.org/ [accessed 2020-02-03]
15. NextGen Healthcare. NextGen® Connect Integration Engine. URL: https://www.nextgen.com/products-and-services/integration-engine [accessed 2020-02-03]
16. What is openEHR? OpenEHR - Open industry specifications, models and software for e-health. Heard S, Beale T. URL: https://www.openehr.org/about/what_is_openehr [accessed 2020-02-03]
17. Ray D, Gulla U, Gupta M, Dash S. Interoperability and Constituents of Interoperable Systems in Public Sector. Handbook of Research on ICT-Enabled Transformational Government 2009:175-195. [doi: 10.4018/978-1-60566-390-6.ch010]
18. IHE International. Profiles. URL: https://www.ihe.net/Profiles/ [accessed 2020-02-03]
19. Website. URL: https://www.w3.org/TR/rdf-sparql-query [accessed 2020-02-03]
20. Craig E. Routledge Encyclopedia of Philosophy. London: Routledge; 1998.
21. OWL - Semantic Web Standards. OWL. URL: https://www.w3.org/OWL/ [accessed 2020-02-03]
22. Programa de Governo Eletrônico Brasileiro. Padrões de Interoperabilidade de Governo Eletrônico. 2018. URL: http://eping.governoeletronico.gov.br/ [accessed 2020-02-03]
23. Crepaldi N, Orfão N, Yoshiura V, Villa T, Netto A, Alves D. Desenvolvimento e implantação de um sistema para a gestão da informação do acompanhamento de doentes de tuberculose. Revista da Faculdade de Medicina de Ribeirão Preto e do Hospital das Clínicas da FMRPUSP 2017:13-17. [doi: 10.11606/d.22.2017.tde-09012017-153505]

24. Lopes P, Oliveira J. A semantic web application framework for health systems interoperability. 2011 Presented at: Proceedings of the first international workshop on Managing interoperability and complexity in health systems - MIXHS 11; 2011; Glasgow, Scotland. [doi: 10.1145/2064747.2064768]
25. Valle E, Cerizza D, Bicer V, Kabak Y, Laleci G, Lausen H. The Need for Semantic Web Service in the eHealth. STI Innsbruck. URL: https://www.w3.org/2005/04/FSWS/Submissions/46/SWS4HC.pdf [accessed 2020-02-03]
26. Abhishek K, Singh MP. An Ontology based Decision support for Tuberculosis Management and Control in India. IJET 2016 Dec 31;8(6):2860-2877. [doi: 10.21817/ijet/2016/v8i6/160806247]
27. Hitzler P, Janowicz K. Semantic Web ? Interoperability, Usability, Applicability. Semantic Web. (1,2) 2010;1:1-2. [doi: 10.3233/sw-2010-0017]
28. Ogundele OA, Moodley D, Pillay A, Seebregts C. An ontology for factors affecting tuberculosis treatment adherence behavior in sub-Saharan Africa. PPA 2016 Apr:669. [doi: 10.2147/ppa.s96241]
29. Gonçalves B, Guizzardi G, Pereira Filho JG. Using an ECG reference ontology for semantic interoperability of ECG data. Journal of Biomedical Informatics 2011 Feb;44(1):126-136. [doi: 10.1016/j.jbi.2010.08.007]
30. V KK. Semantic Web Approach Towards Interoperability and Privacy Issues in Social Networks. IJWSC 2014 Sep 30;5(3):13-17. [doi: 10.5121/ijwsc.2014.5302]
31. Zenuni X, Raufi B, Ismaili F, Ajdari J. State of the Art of Semantic Web for Healthcare. Procedia - Social and Behavioral Sciences 2015 Jul;195:1990-1998. [doi: 10.1016/j.sbspro.2015.06.213]
32. Belleau F, Nolin M, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics 2008 Oct;41(5):706-716. [doi: 10.1016/j.jbi.2008.03.004]
33. McMurray J, Zhu L, McKillop I, Chen H. Ontological modeling of electronic health information exchange. Journal of Biomedical Informatics 2015 Aug;56:169-178. [doi: 10.1016/j.jbi.2015.05.020]
34. Jiang G, Solbrig H, Chute C. Using semantic web technology to support ICD-11 textual definitions authoring. 2012 Presented at: Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences - SWAT4LS 11; 2012; London, United Kingdom. [doi: 10.1145/2166896.2166910]
35. D2R Server ? Publishing Relational Databases on the Semantic Web. Bizer C, Cyganiak R. URL: https://pdfs.semanticscholar.org/afd5/68bbc4d8c5212e13bfa2541296085c5ed45b.pdf [accessed 2020-02-03]
36. The D2RQ Platform ? Accessing Relational Databases as Virtual RDF Graphs. D2RQ. URL: http://d2rq.org/ [accessed 2020-02-03]
37. Pellison FC, Lopes Rijo RPC, Lima VC, de Lima RR, Martinho R, Cruz Correia RJ, et al. Development and evaluation of an interoperable system based on the semantic web to enhance the management of patients' tuberculosis data. Procedia Computer Science 2017;121:791-796. [doi: 10.1016/j.procs.2017.11.102]
38. Filho C, Dias T, Alves D. Arquétipos OpenEHR nas fichas do fluxo do controle da tuberculose. Revista da Faculdade de Medicina de Ribeirão Preto e do Hospital das Clínicas da FMRPUSP. URL: http://revista.fmrp.usp.br/2013/suplementos/revista_IASIS2014.pdf [accessed 2020-02-03]
39. Bizer C, Eckert K, Meusel R, Mühleisen H, Schuhmacher M, Völker J. Deployment of RDFa, Microdata, and Microformats on the Web ? A Quantitative Analysis. Lecture Notes in Computer Science The Semantic Web ? ISWC 2013:2013-2032. [doi: 10.1007/978-3-642-41338-4_2]
40. W3C. HTML Microdata. URL: https://www.w3.org/TR/microdata/ [accessed 2020-02-03]
41. Plastiras P, O'Sullivan DM. Combining Ontologies and Open Standards to Derive a Middle Layer Information Model for Interoperability of Personal and Electronic Health Records. J Med Syst 2017 Oct 28;41(12). [doi: 10.1007/s10916-017-0838-9]
42. Pellison FC, Lima VC, Lopes Rijo RPC, Alves D. Integrating Tuberculosis data in State of São Paulo over Semantic Web: a proof of concept. Procedia Computer Science 2019;164:686-691. [doi: 10.1016/j.procs.2019.12.236]
43. Pahl C, Zare M, Nilashi M, de Faria Borges MA, Weingaertner D, Detschew V, et al. Role of OpenEHR as an open source solution for the regional modelling of patient data in obstetrics. Journal of Biomedical Informatics 2015 Jun;55:174-187. [doi: 10.1016/j.jbi.2015.04.004]
44. Javed M, Abgaz YM, Pahl C. Ontology Change Management and Identification of Change Patterns. J Data Semant 2013 May 22;2(2-3):119-143. [doi: 10.1007/s13740-013-0024-2]
45. Abend O, Yerushalmi S, Rappoport A. UCCAApp: Web-application for Syntactic and Semantic Phrase-based Annotation. 2017 Presented at: Proceedings of ACL , System Demonstrations; 2017; Vancouver, Canada. [doi: 10.18653/v1/p17-4019]
46. Vargas-Vera M, Motta E, Domingue J, Lanzoni M, Stutt A, Ciravegna F. MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web Lecture Notes in Computer Science 2002:379-391. [doi: 10.1007/3-540-45810-7_34]
47. Dill S, Tomlin J, Zien J. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. 2003 Presented at: Proceedings of the twelfth international conference on World Wide Web - WWW 03; 2003; New York, United States of America. [doi: 10.1145/775152.775178]
48. Centro Nacional de Terminologias em Saúde. Accessed February 3, 2020. Centro Nacional de terminologias em saúde: planejamento estratégico 2018 ? 2021. URL: http://portalarquivos2.saude.gov.br/images/pdf/2018/junho/14/planejamento-estrategico-centerms.pdf [accessed 2020-06-10]

49. Rijo R, Martinho R, Oliveira A, Alves D, Reis Z, Santos-Pereira C, et al. Profiling IT security and interoperability in brazilian health organisations from a business perspective. International Journal of E-Health and Medical Communications 2020:2020. [doi: 10.4018/IJEHMC.2020040106]

Abbreviations

BFO: Basic Formal Ontology.
CPF: Cadastro de Pessoas Físicas.
DOTS: directly observed treatment, short-course.
e-SUS: electronic Sistema Único de Saúde.
EHR: electronic health record.
FNI: Formulário de Notificação Individual.
GAL: sistema Gerenciador de Ambiente Laboratorial.
IHE: Integrating the Healthcare Enterprise.
RDF: Resource Description Framework.
SINAN: Notification of Injury Information System.
TB: tuberculosis.
TBWEB: Notification and Monitoring System for Cases of Tuberculosis in the State of São Paulo.

Edited by G Eysenbach; submitted 24.11.19; peer-reviewed by R Martinho, V Della Mea; comments to author 31.12.19; revised version received 17.02.20; accepted 22.03.20; published 06.07.20.

Please cite as:
Pellison FC, Rijo RPCL, Lima VC, Crepaldi NY, Bernardi FA, Galliez RM, Kritski A, Abhishek K, Alves D
Data Integration in the Brazilian Public Health System for Tuberculosis: Use of the Semantic Web to Establish Interoperability
JMIR Med Inform 2020;8(7):e17176
URL: https://medinform.jmir.org/2020/7/e17176
doi:10.2196/17176
PMID:32628611

Original Paper

Prediction of Medical Concepts in Electronic Health Records: Similar Patient Analysis

Nhat Le[1], PhD; Matthew Wiley[1], PhD; Antonio Loza[2], BSc; Vagelis Hristidis[1], PhD; Robert El-Kareh[3], MD, MS, MPH

[1]Department of Computer Science & Engineering, University of California, Riverside, Riverside, CA, United States

[2]School of Medicine, University of California, Riverside, Riverside, CA, United States

[3]Department of Medicine, University of California, San Diego, San Diego, CA, United States

Corresponding Author:
Nhat Le, PhD
Department of Computer Science & Engineering
University of California, Riverside
Winston Chung Hall 363
900 University Ave.
Riverside, CA, 92521
United States
Phone: 1 9518275639
Email: nle020@ucr.edu

Abstract

Background: Medicine 2.0—the adoption of Web 2.0 technologies such as social networks in health care—creates the need for apps that can find other patients with similar experiences and health conditions based on a patient's electronic health record (EHR). Concurrently, there is an increasing number of longitudinal EHR data sets with rich information, which are essential to fulfill this need.

Objective: This study aimed to evaluate the hypothesis that we can leverage similar EHRs to predict possible future medical concepts (eg, disorders) from a patient's EHR.

Methods: We represented patients' EHRs using time-based prefixes and suffixes, where each prefix or suffix is a set of medical concepts from a medical ontology. We compared the prefixes of other patients in the collection with the state of the current patient using various interpatient distance measures. The set of similar prefixes yields a set of suffixes, which we used to determine probable future concepts for the current patient's EHR.

Results: We evaluated our methods on the Multiparameter Intelligent Monitoring in Intensive Care II data set of patients, where we achieved precision up to 56.1% and recall up to 69.5%. For a limited set of clinically interesting concepts, specifically a set of procedures, we found that 86.9% (353/406) of the true-positives are clinically useful, that is, these procedures were actually performed later on the patient, and only 4.7% (19/406) of true-positives were completely irrelevant.

Conclusions: These initial results indicate that predicting patients' future medical concepts is feasible. Effectively predicting medical concepts can have several applications, such as managing resources in a hospital.

(JMIR Med Inform 2020;8(7):e16008) doi:10.2196/16008

KEYWORDS

consumer health information; decision support techniques; electronic health record

Introduction

Background

Medicine 2.0—the intersection of Web 2.0 and health care services, apps, and tools—brings new opportunities for patients to actively contribute to their own care [1]. With the rapid adoption of patients' electronic health records (EHRs) [2], allowing users to find patients with similar experiences and health conditions based on their EHR has the potential to improve the quality of care and expand options for health care solutions [3]. This approach may lead to novel apps for patients, such as self-management recommendations based on big data aggregation across cohorts [4]. Apps that allow patients to find, discuss, and share health data and information can improve

patient outcomes while raising meaningful discussions in disease management [5]. Therefore, finding patients with similar experiences and health conditions is a critical step for patients to contribute to their own care. This capability is becoming more important as more patient records become available (with user consent and commonly anonymized), for instance, through health social networks that aim to connect patients, which drive the need for patient-centered health informatics [6,7].

We evaluated the *hypothesis* that we can predict possible future medical concepts in a patient's EHR by leveraging the EHRs of other patients in the collection. Medical concepts are entities of a medical ontology, which is a knowledge network of medical concepts, where concepts and their definitions are categorized and interconnected (normally via a hierarchy) to present their semantic meanings. Given a point of time, a patient's current medical history is stored in form of EHRs. Future medical concepts are defined as the ones appearing in the patient's EHRs after that point, which is also the patient's future medical record. To evaluate our hypothesis, we first organized each patient's EHR in the database as a list of chronological medical events, which can be divided into a prefix (a sequence of events up to a time moment) and a suffix (a sequence of events that happened after this time moment). Then, we used various interpatient similarity measures to locate other patients' EHRs that have prefixes similar to the current patient's EHR. Finally, we processed the time-based suffixes of the matched EHRs to determine which medical concepts are probable for the future of the current patient's EHR. In short, our method uses EHRs of patients with similar past medical developments to predict a patient's upcoming developments.

Furthermore, our method offers the prediction's explanation by providing similar patients and medical concepts influencing the prediction; thus, it does not suffer the interpretability limitation of common deep learning techniques [8]. Although we used the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II database to evaluate our methods, our methods are applicable to any database of EHRs, where a set of medical concepts can be extracted for various time instances (eg, hospital visits) during a patient's care.

Patients are not the only stakeholders who stand to benefit from the prediction of future medical concepts in an EHR; clinicians and clinical researchers can also benefit from a what-if analysis based on similar patients. For example, when a physician is answering questions for a patient or the patient's family, such an analysis may be helpful as supporting evidence, especially to provide data-driven guidance in the absence of specific gold standard [7]. Moreover, the clinician may view the changes in the probable future EHR of a patient if a specific therapy is undertaken. From a research standpoint, clinical researchers may be interested in finding patients with similar predicted concepts when performing nonrandomized studies, for example, for matching cases and controls.

Related Work

Research related to our study is divided into 2 groups: those that consider (1) interpatient similarity measures and (2) analysis and prediction via aggregated patient data. The former is related to patients with similar experiences, and health conditions were used for predicting future medical concepts. The latter group is related in that an aggregate of patient data across a database of EHRs was used for predicting future medical concepts. However, none of the related studies have defined the notion of EHR prefixes and EHR suffixes when aggregating patient data or finding patients with similar experiences and health conditions.

Interpatient Similarity Measures

When measuring patients with similar experiences and health conditions, we leveraged previous papers, which have studied several interpatient distance functions. Methods include case-based reasoning, vector space models, bag-of-concepts (BoCs), information content, path length between concepts, common ancestors of concepts, and combinations of these. None of these methods have been applied to EHR prefixes and EHR suffixes for predicting future medical concepts. Thus, the intuitive question is, "Are these interpatient similarity measures powerful enough to identify patients with similar histories and futures?"

Cao et al [9] used case-based reasoning to find patients with similar experiences and health conditions based on clinical text. They found that medical concepts are superior features compared with a bag-of-words approach. Similar to this study, the authors restricted medical concepts to a specific subset of semantic types, but the authors did not consider semantic similarity between concepts—for example, 2 concepts may be neighbors in the Systemized Nomenclature of MEDical Clinical Terms (SNOMED-CT) ontology—when comparing patients. Mabotuwana et al [10] studied an ontology-based similarity measure for radiology reports where the authors extended cosine similarity to include the semantic similarity of medical concepts mentioned in radiology reports. The authors found that the addition of semantic similarity allows a vector space model to differentiate between radiology reports of different anatomical and image procedure–based classes. Plaza and Diaz [11] studied concept graphs for measuring interpatient similarity. Given a set of concepts for a patient, all ancestors of each concept are retrieved and assigned a weight based on their depth, where deeper concepts have higher weights. This method is studied in this study and explained in greater detail in the Methods section. Melton et al [12] studied a variety of interpatient distance measures, including BoCs and average path length (APL). Both the BoCs and unweighted APLs are investigated and described in greater detail in the Methods section.

Analysis and Prediction of Aggregated Patient Data

Related work on aggregating patient data for analytics employs a patient database to provide recommendations, analysis, and/or predictions. Gotz et al at IBM Corporation [13-15] developed an interactive system to aid domain experts in retrospective patient cohort analysis. Similar to our study, their system finds a cohort of patients with similar health conditions based on the EHR of the physician's current patient via symptoms. Statistics for the cohort are aggregated and visualized using a variety of techniques, including an outflow graph that models the evolution of symptoms over time and the respective outcomes. Unlike this study, their system does not predict future medical concepts, nor do they use ontologies when measuring patients with similar

health conditions. However, their study complements our study in that the user can use predicted symptoms to explore possible outcomes in the outflow graph.

PatientsLikeMe has also examined the effects of aggregating patient data [4,16]. A web-based survey found that users reported several benefits from having access to aggregated patient statistics. Furthermore, they found a correlation between perceived benefit and the number of website features used by a user, along with demographic similarities between the users of the web-based platform and actual patient populations. This study aimed to complement the data created by PatientsLikeMe by employing aggregated data to predict future medical concepts.

Recent advancements in deep learning offer a new, powerful predictive tool for patients' EHRs [17]. Miotto et al [18] proposed a 3-layered stack of denoising autoencoders to learn a vector representation of each patient from an EHR database of approximately 700,000 patients and then used this *deep patient* embedding to predict the probability of patients developing 78 diseases. Studies by Razavian et al, Lipton et al, Choi et al, and Nguyen et al [19-22] explored the temporal order of medical events and different neural network architectures, such as recurrent convolutional networks. Rajkomar et al [23] represented a patient's entire EHR as a temporal sequence of medical events in the fast health care interoperability resources format and applied various deep learning models to learn the patient's representation for further predictions: inpatient mortality, 30-day unplanned readmission, long length of stay, and 14,025 International Classification of Diseases-9th revision, diagnosis codes. In general, these methods learn the patient's vector representation, which is used to model downstream prediction tasks such as classification or regression problems. Although these studies restrict their predictions to a predefined medical concept set, our study makes predictions of any medical concepts appearing in patients with similar health conditions. Moreover, whereas deep learning approaches offer limited interpretability [8], our method explains how a prediction is made.

Methods

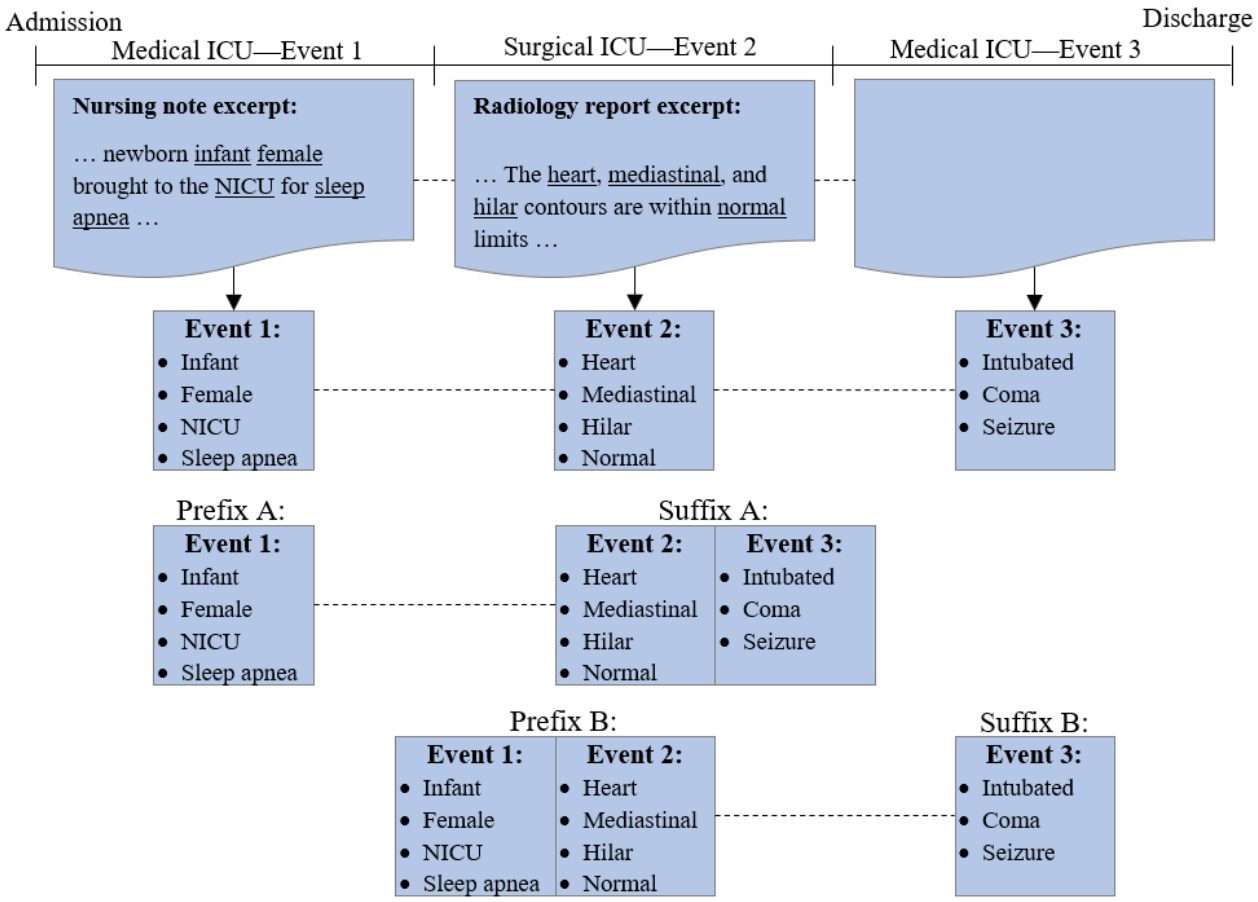
We represented each patient as a set of medical concepts from SNOMED-CT [24]. We extracted medical concepts using the MetaMap library [25]. Then, to identify patients with similar health conditions, we adopted various distance functions studied in the literature [11,12]. We showed how to extend these distance functions to predict future medical concepts, given a query patient. We demonstrated and evaluated these methods on the MIMIC II clinical database, which contains patient data from visits to an intensive care unit (ICU) [26].

Framework and Method for Predicting Future Concepts Using Similar Patients

First, we proposed our framework for discretizing EHRs into events, yielding the notion of *EHR prefixes* and *EHR suffixes*. Consider a database of patient visits to an ICU. One possible method to discretize these visits is to exploit transfers between wards within the ICU, as illustrated by the example in Figure 1. In this example, the patient is admitted to the medical ICU, transferred to the surgical ICU, and then transferred back to the medical ICU. The patient's time in each ward represents a distinct *event*, where clinical notes are recorded that report the patient's status; thus, medical concepts reported in each ward are associated with a specific event. Furthermore, these events have a natural ordering, which produces the notion of EHR prefixes and EHR suffixes. In this example, there are 2 possible EHR prefixes, *[Event1]* and *[Event1, Event2]*, and 2 possible EHR suffixes, *[Event2, Event3]* and *[Event3]*. Hence, each EHR prefix and EHR suffix is associated with a set of medical concepts, as shown at the bottom of Figure 1.

The motivation for discretizing EHRs into events is that health care changes over time with respect to medical conditions, procedures, findings, and drugs observed from the past. Given a new patient, our goal is to find similar EHR prefixes from the EHR database such that the respective EHR suffixes will predict the new patient's future. Let the new patient's EHR be denoted by Q, where Q is represented as a set of medical concepts defined on an ontology. Let Q^p_k represent the set of medical concepts obtained from the first k events, where the superscript p denotes that this set is an EHR prefix. The corresponding EHR suffix is denoted by Q^S_{k+1}, which represents the set of medical concepts from event $k+1$ to the last event in the EHR. Note that in a clinical setting, we would use the whole EHR as Q^p_k as the goal is to predict future concepts, given the current state of the patient. Finally, let D be the database of records within the EHR. We now define our concept prediction algorithm that consists of 2 steps: (1) finding similar records and (2) returning concepts with high confidence.

Figure 1. An example of a patient visiting the intensive care unit, discretized by ward transfers. In this example, the patient was admitted to the medical intensive care unit, transported to radiology, and transferred to the surgical intensive care unit. As this example contains 3 events, there are 2 possible electronic health record prefixes and 2 possible electronic health record suffixes. ICU: intensive care unit; NICU: neonatal intensive care unit.



Concept Prediction Algorithm

Step 1: Compute Similar Electronic Health Record Prefixes

In particular, find the set S of EHR suffixes that correspond to the EHR prefixes P_i in D whose dissimilarity with respect to Q^p_k is less than some *dissimilarity threshold* τ: where P_i is an EHR prefix of events from a single visit, S_i is the corresponding EHR suffix, and *DisSim* is an interpatient dissimilarity function. Note that we only considered the most similar EHR prefix for each visit.

Step 2: Return Concepts With High Confidence

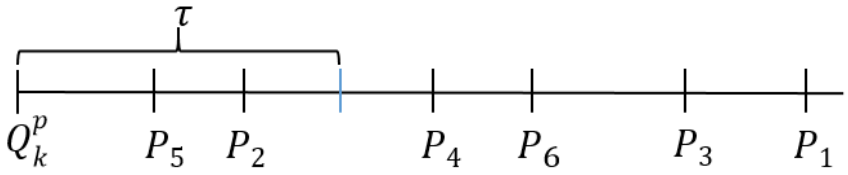
Let be the confidence of concept c, where S'_c is the EHR suffixes from S that contain c. We return , which is the set of concepts in S with confidence greater than the *confidence threshold* .

Figure 2 illustrates step 1 of the concept prediction algorithm, where only prefixes P_2 and P_5 have dissimilarities from the query prefix p (or with respect to Q^p_k) smaller than the threshold τ; thus, their corresponding suffixes S_2 and S_5 are included in S. Define . Furthermore, let P_5 and S_5 be EHR prefix B and EHR suffix B from Figure 1. Thus,

.

Let λ=0.7, then step 2 of the algorithm returns C={*Intubated*, *Seizure*}.

Figure 2. Dissimilarities of electronic health record prefixes with respect to the k-events prefix of a patient Q denoted by Q_k^p.

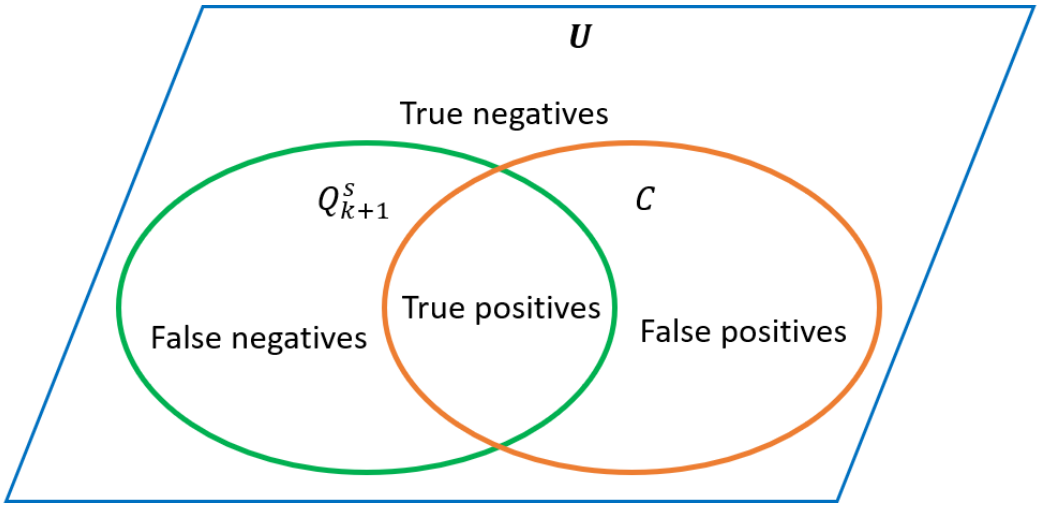


Hence, we can evaluate both parameters and *DisSim* using traditional measures of specificity, sensitivity, and precision. Let, U be the universe of all medical concepts. True-positives (TPs), true-negatives (TNs), false-positives (FPs), and false-negatives (FNs) are defined by:

We have also extended our definitions of TP, TN, FP, and FN to consider *fresh concepts* only. Fresh concepts are concepts that appear in the query EHR suffix,  , which do not appear in the query EHR-prefix,  . We argue that fresh concepts are more challenging and have a higher potential to be clinically useful for prediction. We analyzed fresh concepts separately from all concepts as concepts that appear in the query EHR prefix are likely to persist into the suffix and thus would skew our evaluation of fresh concepts. Therefore, we ignore concepts that appear in  when evaluating any measures concerning TP, TN, FP, or FN.

Figure 3 illustrates the connection between the entire set of concepts *U*, the predicted set of concepts *C*, and the ground truth  . In our experiments, the size of  , and thus, the number of TNs skews the value of specificity. Therefore, we assessed the parameters and interpatient distance measures using the harmonic mean of sensitivity and precision, commonly known as the F-measure in information retrieval.

Figure 3. The connection between the ground truth concepts and the predicted concept space.



Interpatient Distance Measures

We evaluated 4 interpatient dissimilarity measures proposed in the literature [4,5]: (1) *BoC*, (2) *CAs*, (3) *APL*, and (4) symmetric APL (*APL_SYM*).

Let *A* and *B* be the sets of medical concepts.

For BoC, the dissimilarity between *A* and *B* is defined as the sum of the number of concepts that appear in A but not in B and in B but not in A, divided by the size of their union [5]. union of *A* and *B* is also a set, and therefore, the size of the union only considers each concept once:

BoC produces values between 0 and 1, where 0 represents maximum similarity, and 1 represents minimum similarity. Note that BoC is symmetric; hence, *BoC*(*A, B*)=*BoC*(*B, A*).

In CA, for each concept, for each concept c_a in A, we retrieved all ancestor concepts in the concept hierarchy and assigned to each concept and its ancestors a weight, where each c_a is assigned a weight of 1, and ancestors of each c_a are assigned a weight relative to their distance from c_a. An analogous weighting procedure is applied to all concepts and their ancestors in *B*. Weights are averaged if a node is assigned more than one weight.

Let *A′* and *B′* be the set of concepts and their ancestors for *A* and *B*, respectively. When computing the dissimilarity from *A* to *B*, we examined each concept in *A'* and check if it exists in *B'*. If it exists, the given concept in *A'* is assigned a value equal to its own weight, and zero otherwise [4]:

where $w(c_i)$ is the weight assigned to the concept c_i. Hence, the abovementioned sum measures the overlap between the concepts and the ancestors of *A* and *B*. Scores from CA range from 0 to 1, where a score of 0 represents maximum similarity, and 1 represents minimum similarity. By definition, CA is not symmetric.

The APL measure finds the minimum number of edges between each concept in A with every concept in B. APL sums the distances across all concepts in A to obtain the dissimilarity of *A* to *B* [5]:

A score of 0 implies a maximum similarity. By definition, APL is not symmetric; APL_SYM is the sum of *A* to *B* and *B* to *A*:

Preparation of Multiparameter Intelligent Monitoring in Intensive Care II Data Set

We applied our framework and the aforementioned interpatient dissimilarity measures to the MIMIC II clinical database—a database of EHRs collected over a 7-year period from multiple ICUs at a medical center in Boston [26]. Several types of clinical notes are recorded during a visit, including radiology reports,

nursing notes, and physician notes. We parsed each note to extract medical concepts from the clinical text. Each note is associated with a timestamp that represents its creation time. We used these timestamps to map notes to events, defined as ward transfers, generating a list of concepts for each event.

First, we parsed medical concepts from each type of note using the MetaMap library [25]. Before parsing each note, abbreviations such as *OMG* were identified and expanded using an abbreviation list similar to the list of Wiley et al [27]. The MetaMap library maps free text to biomedical concepts are defined in the Unified Medical Language System (UMLS) [28]. Each concept in the UMLS corresponds to one or more semantic types [29], which further maps to semantic groups [30]. Previous studies have shown that disorders, physiology, chemicals and drugs, procedures, and anatomy are the most important UMLS semantic groups when measuring interpatient similarity [11]. Negated concepts are identified via MetaMap, and these concepts are ignored, as previous work has shown that absent concepts are not relevant to patient similarity [11]. After obtaining a list of relevant concepts, each concept from the UMLS is converted to a concept from SNOMED-CT using the MRCONSO table [31].

A single patient visit may consist of several transfers between wards. Each of these transfers is considered to be a *census event* in the MIMIC II database. The rationale for this definition of an event is that each time a patient enters a new care unit, there may be a significant change in the patient's status, for example, the patient's condition worsened, and he was transferred to the surgical ICU.

If a patient visits a hospital multiple times, each visit is treated independently, that is, multiple visits are viewed as different patients for the purpose of our similarity matching algorithm. This decision is not critical for the MIMIC II data set because a majority of patients only have one visit. Related work has shown that the abovementioned concept of census events provides an effective timeline of a patient's record, where concepts within an event are semantically associated with each other [32].

Computation Time Analysis

The computation cost to extract ancestors is linear with respect to the number of ancestors. As the ontology is a wide directed acyclic graph (DAG) instead of a deep one, each concept has up to 61 ancestors, and 29 ancestors on average. We used Dewey encoding to speed up both the retrieval of ancestors and calculation of concept distance. In particular, a concept's Dewey encoding encapsulates its ancestor information, for example, if concept *C2315591* is encoded as *$.8.96.45*, this implies that the concept's ancestors are *$.8* and *$.8.96*. Using Dewey [33] encodings, the distance between 2 concepts is reduced to be a string comparison between their encodings; that is, we computed the distance between the concepts and their lowest common ancestor, which again has cost linear on the DAG depth.

Results

Anecdotal Example

We started with a real anonymized example from the MIMIC II dataset to demonstrate the potential utility of our approach. Bob was involved in a motor vehicle collision where he struck his head and lost consciousness. He arrived at the medical ICU with a chief complaint of severe shoulder pain and bleeding from his nostrils. After arriving at the medical ICU (event 1), Bob was transferred to the surgical ICU for further care (event 2). During his stay in the surgical ICU, the staff observed symptoms of pneumonia and pulmonary aspiration. Bob was then transported to radiology (event 3), where tests revealed that Bob indeed had both pneumonia and pulmonary aspiration. We executed our prediction method using event 1 as a query. In particular, we used *CA*, with τ=0.5 and λ=0.3. Of the suffixes of patients with similar EHR prefixes, 50% contain the concepts of pneumonia and pulmonary aspiration, whereas 29% and 23% of all patients in the general ICU population contained the concepts of pneumonia and pulmonary aspiration, respectively.

Event-Based Analysis of the Multiparameter Intelligent Monitoring in Intensive Care II Data Set

We only considered visits with more than one event because visits with 1 event cannot be split into EHR prefixes and EHR suffixes. In total, there are 4083 visits over 3971 unique patients; thus, patients with multiple visits account for less than 3% of the total number of visits. Visits with 2 events dominate the data set, accounting for 80% of the total visits, whereas visits with 3 events accounted for 15% of the total visits. In general, a longer visit produces more medical concepts, implying that new concepts are found as the patient's visit progresses. Visits of length 2, 3, and 4, respectively, have 291, 434, and 539 unique medical concepts on average. The corresponding number for visits of more than 4 events is 725. On average, each event contains 187 medical concepts, and each visit contains 325 medical concepts. Furthermore, these concepts are dominated by disorders (36%) and procedures (22%). The other concept semantic groups are anatomy (20%), drugs (12%), and physiology (10%).

Prediction Results

We evaluated the interpatient distance measures BoC, CA, APL, and APL_SYM on the aforementioned admissions of the MIMIC II database using our framework of EHR prefixes and EHR suffixes. Our first objective was to tune the parameters τ and λ using the *F* measure. We split the admissions into training and testing datasets, where 20% of the admissions were used for training, and 80% of the admissions were used for testing. Table 1 reports the combination of τ and λ that produced the highest *F* measure for each interpatient distance measure using the training data set. APL_SYM obtains the highest *F* measure, precision, and sensitivity, whereas APL obtains the highest specificity.

Table 1. The best parameters for each distance function based on the training data set.

DisSim	τ	λ	*F* measure (%)	Specificity (%)	Sensitivity (%)	Precision (%)
Bag-of-concept	0.7	0.08	51.8	86.9	52.9	50.6
Common ancestor	0.46	0.25	48.9	94.0	55.2	43.9
Average path length	1.5	0.30	48.7	94.9	52.6	45.4
Symmetric average path length	1.86	0.07	52.4	84.4	52.9	52.0

Figure 4 illustrates a graphical representation of the optimal parameters reported in Table 1, plotting λ on the y-axis and 1−τ on the x-axis. Thus, all concepts from the EHR suffixes of similar EHR prefixes are included with a score to the right of the corresponding vertical dashed line, and from these concepts, all concepts with a confidence above the corresponding horizontal dashed line are included in the predicted EHR suffix. Furthermore, APL and APL_SYM have been normalized by the maximum possible similarity score, where the maximum similarity score is defined as the maximum path length in SNOMED-CT. As shown in this figure, CA and BoC have larger values of dissimilarity compared with APL and APL_SYM. The tightest bounds for both thresholds are for APL and APL_SYM, and the loosest bound is for BoC. This is expected, as the average scores for BoC, CA, APL, and APL_SYM are 0.86, 0.31, 0.07, and 0.07, respectively. Moreover, APL and CA have tightest bounds on the confidence threshold; this is an interesting point, as APL and CA are antisymmetric, implying that *symmetric interpatient distance measures require less confidence when predicting future medical concepts*.

Table 2 reports the results on the testing dataset using the optimal set of parameters reported in Table 1 for fresh and not fresh concepts.

Figure 4. Representation of the optimal choice of the dissimilarity threshold τ and confidence threshold λ for the training data set. APL: average path length; BoC: bag-of-concept; CA: common ancestor; APL-SYM: symmetric average path length.

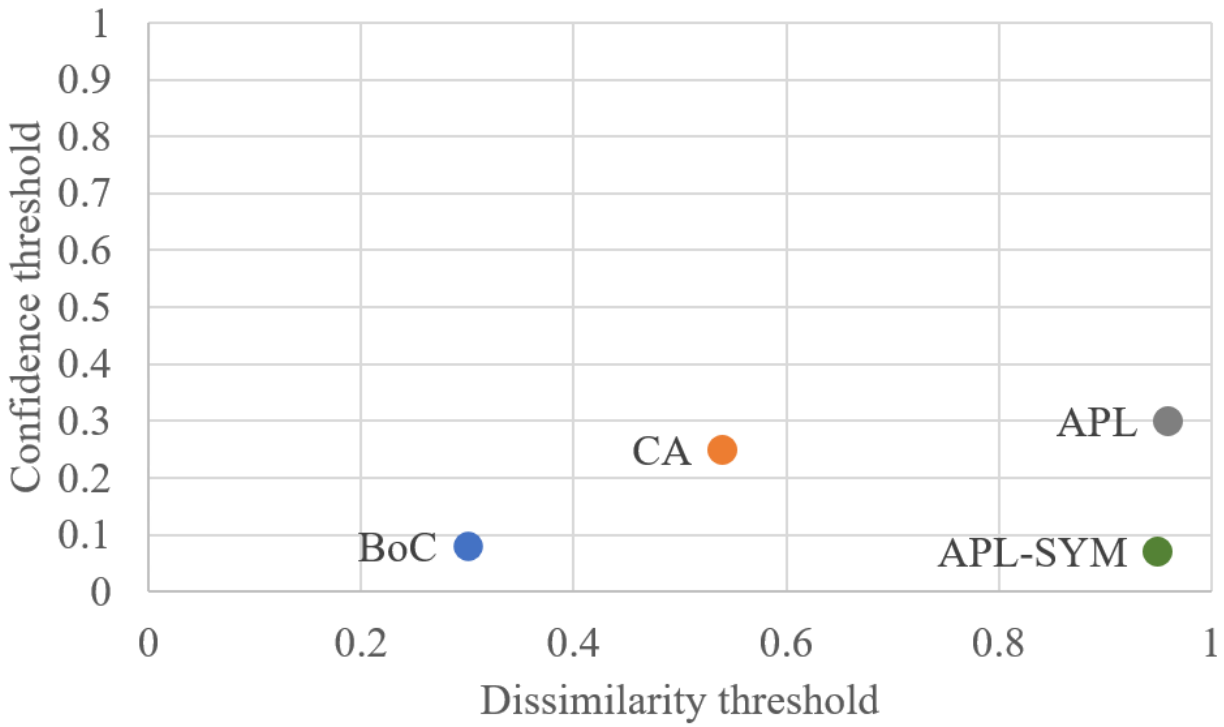


Table 2. The results for the testing data set separated by semantic group, using the parameters tuned on the training data set for fresh and not fresh concepts.

Semantic group and *DisSim*	*F* measure (%)	Specificity (%)	Sensitivity (%)	Precision (%)
All concepts				
BoC[a]	51.7	87.1	52.6	50.8
CA[b]	48.9	94.1	55.7	43.7
APL[c]	48.7	94.8	52.8	45.2
APL_SYM[d]	*52.3* [e]	84.2	53.7	50.9
Disorders				
BoC	49.4	87.8	49.8	49.0
CA	44.7	95.0	48.4	41.4
APL	44.4	95.7	46.1	42.8
APL_SYM	*50.7*	85.0	51.8	49.6
Procedures				
BoC	52.2	85.6	53.3	51.3
CA	48.9	92.6	57.9	42.3
APL	48.0	93.6	54.0	43.2
APL_SYM	*52.6*	82.0	54.4	51.0
Chemicals and drugs				
BoC	49.7	89.8	49.1	50.4
CA	48.5	96.4	49.9	47.1
APL	48.1	96.9	47.3	48.9
APL_SYM	*50.2*	87.7	49.8	50.7
Physiology				
BoC	56.6	82.3	57.1	56.1
CA	57.8	89.5	69.5	49.5
APL	*58.4*	90.6	67.6	51.4
APL_SYM	56.9	80.1	57.7	56.1

[a]BOC: bag-of-concept.

[b]CA: common ancestor.

[c]APL: average path length.

[d]APL_SYM: symmetric average path length.

[e]Italicized numbers indicate the best result of the semantic group.

Similarly, Table 3 reports the same results for fresh concepts only; *fresh concepts are concepts that do not appear in the query EHR prefix and, therefore, are fresh to the query EHR suffix*. We categorized each concept into its semantic group and analyzed each interpatient distance measure with all concepts and concepts restricted to a semantic group; anatomical concepts are omitted in this analysis, as predicting an anatomical site, such as lower back, is not useful in a clinical setting.

Table 3. The results for the testing data set separated by semantic group, using the parameters tuned on the training data set for fresh concepts only.

Semantic group and *DisSim*	*F* measure (%)	Specificity (%)	Sensitivity (%)	Precision (%)
All concepts				
BoC[a]	43.7	89.6	43.3	44.1
CA[b]	34.8	95.7	37.7	32.4
APL[c]	34.5	96.5	35.0	34.0
APL_SYM[d]	*44.9* [e]	86.8	45.3	44.6
Disorders				
BoC	42.1	90.0	40.8	43.5
CA	32.1	96.3	32.9	31.3
APL	31.8	97.0	30.6	33.1
APL_SYM	*44.1*	87.3	43.7	44.5
Procedures				
BoC	43.6	88.5	42.7	44.5
CA	35.6	94.8	39.3	32.5
APL	34.7	95.6	36.4	33.2
APL_SYM	*44.7*	85.0	44.6	44.7
Chemicals and drugs				
BoC	38.6	91.8	34.9	43.1
CA	30.4	97.5	26.6	35.6
APL	28.7	97.9	23.7	36.3
APL_SYM	*39.9*	89.7	36.5	44.1
Physiology				
BoC	46.1	86.1	45.2	47.0
CA	40.9	92.8	45.3	37.3
APL	41.2	93.8	43.4	39.2
APL_SYM	*47.2*	83.9	46.8	47.5

[a]BOC: bag-of-concept.

[b]CA: common ancestor.

[c]APL: average path length.

[d]APL_SYM: symmetric average path length.

[e]Italicized numbers indicate the best result of the semantic group.

As shown in Table 2, the symmetric interpatient distance measures outperform the antisymmetric distance measures across all semantic groups, where APL_SYM performs the best; the only exception is physiology. Comparing these results with Table 3 shows that the gap between symmetric and antisymmetric distance measures widens to a 10% difference in terms of *F* measure. That is, *symmetric interpatient distance measures are more predictive of future medical concepts, especially for fresh concepts*. When considering the symmetric measures APL_SYM and BoC, *APL_SYM consistently performs better*, achieving higher rates of sensitivity and precision in every case.

Furthermore, the *antisymmetric interpatient distance measures performed better with respect to specificity* but achieved a lower precision. That is, antisymmetric distance measures predicted fewer concepts overall to achieve higher rates of specificity with lower rates of sensitivity and precision, which is explained by the conservative choice made during the tuning phase. Another interesting point is that all interpatient distance measures observed an increase in specificity for fresh concepts; however, this increase was greatest for symmetric interpatient distance measures. The reason is that the number of FP decreases for fresh concepts, whereas the nonfresh concepts are more frequently predicted to be in the suffix and, therefore, have a higher frequency of FPs.

Clinical Significance of the Subset of Predicted Concepts

We further examined 16 individual concepts identified as important by our physician author (RE) in the ICU setting. We focused on the TP cases (correctly predicted mention in the suffix) to validate the prediction's importance and FN cases

(incorrectly predicted no mention in the suffix) to detect possible significant misses. We presented our predictions in a web interface (Table 4), which is basically a table of predicted concepts, the patient's EHR prefix/suffix and concepts influencing the prediction in highlight.

Table 4. Predictions and explanations provided to our medical student and physician authors to label the clinical significance of a prediction.

Patient ID	Predicted concept and time	Prefix at time of prediction	Suffix from time of prediction
22,487	*Bronchoscopy* (3 hours:23 min:0 seconds)	...Resp: RR 16-20 has periods of apnea when asleep... …There is increased density in the right upper lung field with elevation of the minor fissure consistent with developing atelectasis in the right upper lobe…	…*Bronchoscopy* done secondary to low PaO2...

In Table 4, our domain expert is given a prediction, the patient history, and asked to evaluate if the prediction is helpful. Particularly, in the third column (*Prefix at time of prediction*), we presented the patient history up to the point that our system predicts that a concept(s) will appear in future (in the second column *Predicted concept and time*). The last column in Table 4 (*Suffix from time of prediction*) shows events occurring after the prediction time so that our domain expect can judge if the system's prediction is significant in the sense that the predicted concepts actually affect the patient and the prediction is not trivial, that is, obviously happen, thus no need for prediction. As we focused on the positive cases, the predictions actually appear in the patient's suffix and thus are highlighted for the domain expert to evaluate.

Our medical student and physician authors manually mark each case with 1 of 4 categories: (1) mentioned and performed; (2) concept mentioned but it is obvious (ie, little value to clinicians); (3) mentioned but only considered by physician, not performed (ie, the clinicians mentioned this concept in the suffix but in the end did not perform the procedure); and (4) mentioned, but out of context (eg, mentioned as part of the medical history of a patient or while describing a similar case). We reported additional metrics such as specificity, sensitivity, FP, and TN of 7 important concepts in the Multimedia Appendix 1, ordered by concept name. We do not count the cases in which a predicted concept occurs in both the patient's prefix and suffix. Moreover, if a patient history can be divided into multiple prefix-suffix pairs and the algorithm is able to make predictions for a long prefix, not for the shorter prefix, we do not count the case of a shorter prefix as a negative prediction.

True-Positive Analysis

Table 5 reports the fine-grained evaluation of TP cases. Note that we only presented predictions of 7 concepts because our algorithm did not predict the remaining 9 concepts. The bronchoscopy concept was successfully mentioned and performed in the suffix 63 of 63 times in a TP category. Bronchoscopy was positively identified with the keywords in the prefix, usually mentioning respiratory symptoms. Compared with bronchoscopy, surgery is a much more invasive procedure that requires consent of the patient and for the patient to be medically cleared for surgery. This caused 215 surgical concepts to be accurately mentioned and performed but have a significant portion mentioned out of context (16 times) or mentioned but only considered and not performed (25 times). Patients have a craniotomy performed for a variety of reasons. One craniotomy in the medical records analyzed was accurately mentioned and performed, but it was not needed to be predicted. The patient undergoing a craniotomy came in after a motor vehicle collision with an obvious facial fracture, thus not needing to predict the craniotomy, as it would be the only way to treat the patient. In summary, most TP predictions are useful. Overall, 13.1% of the predictions are unhelpful, and mostly fall into the surgery concept.

Table 5. Expert evaluation of true positive predictions using 4 fine-grained categories.

Concept	Mentioned and performed	Concept mentioned, but is obvious	Mentioned but only considered by physician, not performed	Mentioned, but out of context
Bronchoscopy	63	0	0	0
Cardiac surgery	5	0	0	0
Colonoscopy	1	0	0	0
Craniotomy	9	1	1	1
Dialysis procedure	47	0	6	1
Refractive surgery enhancement	13	0	0	1
Surgery	215	1	25	16

We illustrated how our algorithm offers useful predictions using a TP case example. In patient ID 22,487, a bronchoscopy was successfully predicted in the suffix (Table 4). The patient had a history of coronary artery disease with chest pain and had a triple coronary artery bypass graft performed to alleviate his symptoms before the prefix. In the prefix, our algorithm highlighted (we *highlighted* a concept in the prefix if it is contributing to the prediction of the target concept in the suffix) *effusion* 7 times, *apnea* 6 times, and *increased density* one time, all related to pulmonary pathology. Heparin, a blood thinner,

was also highlighted 7 times by our algorithm. The patient' s respiratory state began to diminish and was eventually placed on a ventilator, as his course in the hospital progressed. Bronchoscopy was accurately predicted and performed on day 3 and hour 23 in the suffix *secondary to low PaO_2* with small amounts of suctioned thin secretions, and no plugs were found. The accurately predicted concept is interesting, as the patient was initially presented with chest pain–related symptoms treated by intervention through the cardiovascular organ system but was found to have concurrent complications in the pulmonary organ system.

To obtain the full picture, we presented a TP example that is clinically incorrect. In patient 9122, a surgery was predicted in the suffix, but no performance of a surgery in the suffix was found. This patient was a 25-week premature twin baby born by cesarean section. The only mention of surgery in the suffix is an update by a neonatal intensive care unit nurse stating they were *awaiting surgical time for twin*. No surgery was considered or performed for this patient during the suffix and was only being medically managed for being born prematurely. One of the most highlighted words in the prefix used by the algorithm to predict surgery was *bili* with 35 mentions, *bilirubin* had 3 mentions, and *phototherapy* with 20 mentions—all related to jaundice. There were also multiple highlighted words related to respiratory symptoms, such as *gas* with 18 mentions, *bicarb* having 9 mentions, and 3 mentions for *PCO_2* Although no surgery plan was considered for the patient, the word surgery was present in the suffix, that is, this is an *out of context* prediction.

In Multimedia Appendix 2, we examined how early our algorithm can predict concept occurrences. In particular, in TP cases, we calculated the time from the prefix's end to the suffix's beginning. For most concepts, the minimum times are almost 0 because there are suffixes that occur right after their prefixes. On average, our algorithm can predict concepts several days before their actual occurrences.

False-Negative Analysis

We presented the same evaluation on FN cases in Table 6. Although 53 bronchoscopies were accurately mentioned and performed, the FN had an additional concept mentioned in context (1 time) or mentioned but only considered and not performed (3 times). Colonoscopy appeared more in the FN group with 21 colonoscopies mentioned and performed but had a high quantity of concepts mentioned in context (5) or mentioned but only considered and not performed (13). The surgery group also mentioned and performed 154 concepts; however, similar to Table 5, it has a significant number of predictions made out of context (8) or mentioned but only considered and not performed (42). The refractive surgery enhancement concept had the lowest ratio of concepts accurately mentioned and performed (48) to those mentioned out of context (21) or mentioned but only considered and not performed (14). Overall, 24.8% of FN cases are unimportant because of being out of context or not being performed by physicians.

Table 6. Expert evaluation of false negative predictions using 4 fine-grained categories (for instance, surgery was not predicted to be in suffix, and it appears in the suffix).

Concept	Mentioned and performed	Concept mentioned, but not needed for prediction	Mentioned but only considered by physician, not performed	Mentioned, but out of context
Bronchoscopy	49	0	3	1
Cardiac surgery	40	0	6	0
Colonoscopy	26	0	13	5
Craniotomy	23	0	2	1
Dialysis procedure	46	0	6	1
Refractive surgery enhancement	48	0	14	23
Surgery	154	0	43	9

Discussion

Principal Findings

Our results show that when applied to clinical concept prediction in ICU patients, symmetric interpatient distance measures are more robust in terms of *F* measure, sensitivity, and precision. Furthermore, antisymmetric interpatient distance measures performed the best in terms of specificity. Hence, antisymmetric interpatient distance measures are more conservative when predicting future medical concepts, as explained by their high confidence thresholds and high levels of specificity, whereas symmetric interpatient distance measures observe a 10% gain in precision and sensitivity over antisymmetric measures. Thus, symmetric interpatient distance measures are more predictive of future medical concepts. Overall, the APL_SYM performed the best.

We further evaluated the clinical value of the predictions. Our medical student and physician authors manually examined the TP and FN predictions of 16 important concepts. We found that 86.9% (353/406) of TP predictions are performed later, and only 4.7% (19/406) of the cases are totally out of context. This early concept prediction capability implies substantial impacts, such as avoiding potential high-risk events and improving patient outcomes at lower costs. On the other hand, our algorithm missed 513 FN cases, but 24.7% of them were clinically unimportant. Specifically, these missed concepts do appear in the patient suffixes but are out of context, or not needed, or not performed by the physician.

As an example of an application of the proposed methods in a real setting, we considered using these methods to periodically automatically predict the estimated number of patients in a hospital that will require bronchoscopy. This may allow for better resource planning.

Limitations

We recognized that in its current form, our system is not sufficiently accurate for deployment. In particular, concern arises when giving a patient or their family access to our proposed methods—incorrectly predicting an undesired concept may incur unneeded stress and anxiety. In this regard, we may calibrate the confidence parameters to achieve higher precision and have an expert manually select the set of concepts that are appropriate to present to patients. As an example of a potential application, such a controlled prediction module could be deployed in a patient portal of a health insurance company, where a patient can already view his or her EHR.

From a medical perspective, ICUs are often numerically oriented with vital signs, pressure readings, laboratory values, and ventilator readings. Furthermore, ICUs move at a fast pace, and hence, using the granularity of ward transfers is perhaps too broad in the ICU setting. Therefore, our proposed methods will most likely achieve different results in a primary care or outpatient setting. An interesting analysis would be to compare long-term predictions in the outpatient setting with near-term predictions in the ICU setting.

However, the MIMIC database is one of the few, if only publicly available databases of EHRs that are rich in both clinical notes and temporal data. Clinical notes enable a rich collection of clinical concepts and hence allow for the prediction of a broad range of clinical concepts. *For example*, an EHR database containing only disease classifications will represent diabetes but will fail to represent insulin; hence, insulin cannot be predicted. Furthermore, temporal data allow us to sort medical concepts into prefixes and suffixes.

Another medical limitation is that we did not weigh concepts based on their clinical importance. For example, the concept of *cardiac arrest* is more important in terms of similarity and predictive value than the concept of *coughing*. Moreover, the importance of a clinical concept depends on its application and domain. Furthermore, we need to assess the accuracy required for our system to be useful to patients, clinicians, and researchers. This accuracy requirement could be assessed through user evaluations.

From a technical perspective, a key limitation is the assumption that MetaMap correctly identifies all concepts written in a clinical note. MetaMap has achieved reasonable precision and recall values (80% and 79%, respectively) when identifying medical concepts from clinical notes [34]. Given the raw text of a clinical note, this assumption is clearly invalid because of abbreviations in the clinical note and errors generated by MetaMap. We address abbreviations by using a manually crafted list of medical abbreviations common to clinical notes; thus, potential errors caused by ambiguities because of common abbreviations were minimized. Furthermore, we argue that errors generated by MetaMap are a natural language processing problem, which is beyond the scope of this study. MetaMap limitation also holds with any other automatic extraction tool. To mitigate this, our physician author manually evaluated the clinical significance of TP predictions for a subset of interesting concepts.

Another technical limitation is that we evaluated our algorithm strictly, in that we only accepted predictions that exactly predicted the corresponding concept. *For example*, if we predicted *cancer* when the actual concept was *breast cancer*, then our prediction of cancer would be marked as an FP, when our prediction was semantically relevant. Hence, including semantically similar concepts, either through is-a (ISA) ancestors or other semantic relations, has the potential to increase the accuracy of our algorithm while remaining relevant to clinical decision support.

Conclusions

In this paper, we studied the problem of predicting future medical concepts in a patient's EHR. The key idea of our method was to find patients with similar EHR prefixes using various interpatient similarity measures and then predict medical concepts that have high confidence in EHR suffixes of those patients. Our results showed that this is a promising approach to predict possible future concepts in a patient's EHR. Of the multiple symmetric and antisymmetric interpatient similarity measures, the APL_SYM achieved the highest accuracy in our evaluation. We further evaluated the predictions of 16 important concepts manually and found that 86.9% of TP predictions are performed later. These initial results indicate that predicting a patient's future medical concepts is feasible.

Acknowledgments

This project was partially supported by the National Science Foundation grants IIS-1838222, IIS-1619463, and IIS-1901379.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prediction performance results for important concepts selected by our physician author. We do not count the cases that a predicted concept occurs in both patient's prefix and suffix.
[DOCX File , 14 KB - medinform_v8i7e16008_app1.docx]

Multimedia Appendix 2
Time from our algorithm prediction to the actual occurrence of the concepts in suffix for true positive cases (The time is formatted as dd hh:mm:ss, where dd is dropped if the time is less than a day).
[DOCX File , 14 KB - medinform_v8i7e16008_app2.docx]

References

1. van de Belt TH, Engelen LJ, Berben SA, Schoonhoven L. Definition of health 2.0 and medicine 2.0: a systematic review. J Med Internet Res 2010 Jun 11;12(2):e18 [FREE Full text] [doi: 10.2196/jmir.1350] [Medline: 20542857]
2. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81 [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
3. Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. Int J Environ Res Public Health 2009 Feb;6(2):492-525 [FREE Full text] [doi: 10.3390/ijerph6020492] [Medline: 19440396]
4. Wicks P, Keininger DL, Massagli MP, de la Loge C, Brownstein C, Isojärvi J, et al. Perceived benefits of sharing health data between people with epilepsy on an online platform. Epilepsy Behav 2012 Jan;23(1):16-23 [FREE Full text] [doi: 10.1016/j.yebeh.2011.09.026] [Medline: 22099528]
5. Frost JH, Massagli MP. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. J Med Internet Res 2008 May 27;10(3):e15 [FREE Full text] [doi: 10.2196/jmir.1053] [Medline: 18504244]
6. Frost JH, Massagli MP, Wicks P, Heywood J. How the social web supports patient experimentation with a new therapy: the demand for patient-controlled and patient-centered informatics. AMIA Annu Symp Proc 2008 Nov 6:217-221 [FREE Full text] [Medline: 18999176]
7. Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. Health Aff (Millwood) 2014 Jul;33(7):1229-1235. [doi: 10.1377/hlthaff.2014.0099] [Medline: 25006150]
8. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. J Am Med Assoc 2017 Aug 8;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
9. Cao H, Melton GB, Markatou M, Hripcsak G. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. J Biomed Inform 2008 Dec;41(6):882-888 [FREE Full text] [doi: 10.1016/j.jbi.2008.03.006] [Medline: 18487093]
10. Mabotuwana T, Lee MC, Cohen-Solal EV. An ontology-based similarity measure for biomedical data-application to radiology reports. J Biomed Inform 2013 Oct;46(5):857-868 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.013] [Medline: 23850839]
11. Plaza L, Díaz A. Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs. In: Proceedings of the International Conference on Application of Natural Language to Information Systems. 2010 Presented at: NLDB'10; June 23-25, 2010; Cardiff, United Kingdom p. 296-303. [doi: 10.1007/978-3-642-13881-2_31]
12. Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform 2006 Dec;39(6):697-705 [FREE Full text] [doi: 10.1016/j.jbi.2006.01.004] [Medline: 16554186]
13. Wongsuphasawat K, Gotz D. Outflow: Visualizing Patient Flow by Symptoms and Outcome. In: Proceedings of the IEEE VisWeek Workshop on Visual Analytics in Healthcare. 2011 Presented at: IEEE VisWeek'11; October 23, 2011; Providence, RI URL: https://www.semanticscholar.org/paper/Outflow-%3A-Visualizing-Patient-Flow-by-Symptoms-and-Wongsuphasawat-Gotz/f82bc74b05438a6739d51b78e4a64a78fc29a67b
14. Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE Trans Vis Comput Graph 2012 Dec;18(12):2659-2668. [doi: 10.1109/TVCG.2012.225] [Medline: 26357175]
15. Zhang Z, Gotz D, Perer A. A Visual Analysis Approach to Cohort Study of Electronic Patient Records. In: Proceedings of the Conference on Bioinformatics and Biomedicine. 2014 Presented at: BIBM'12; November 2-5, 2014; Seattle, WA. [doi: 10.1109/BIBM.2014.6999214]
16. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res 2010 Jun 14;12(2):e19 [FREE Full text] [doi: 10.2196/jmir.1549] [Medline: 20542858]
17. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
18. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016 May 17;6:26094 [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
19. Razavian N, Marcus J, Sontag D. Multi-Task Prediction of Disease Onsets from Longitudinal Laboratory Tests. In: Proceedings of the 1st Machine Learning for Healthcare Conference. 2016 Presented at: PMLR'16; August 19-20, 2016; Los Angeles, CA p. 73-100 URL: http://proceedings.mlr.press/v56/Razavian16.html

20. Lipton Z, Kale D, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint 2015 epub ahead of print - 1511.03677 [FREE Full text]
21. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In: Proceedings of the Conference on Machine Learning and Healthcare Conference. 2016 Presented at: MLHC'16; August 19-20, 2016; Los Angeles, CA.
22. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: a convolutional net for medical records. IEEE J Biomed Health Inform 2017 Jan;21(1):22-30. [doi: 10.1109/JBHI.2016.2633963] [Medline: 27913366]
23. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
24. Stearns M, Price C, Spackman K, Wang A. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662-666 [FREE Full text] [Medline: 11825268]
25. Aronson A. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
26. Saeed M, Lieu C, Raber G, Mark R. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol 2002;29:641-644. [Medline: 14686455]
27. Wiley MT, Jin C, Hristidis V, Esterling KM. Pharmaceutical drugs chatter on online social networks. J Biomed Inform 2014 Jun;49:245-254 [FREE Full text] [doi: 10.1016/j.jbi.2014.03.006] [Medline: 24637141]
28. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database Issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
29. National Library of Medicine. Current Semantic Types URL: http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html [accessed 2019-08-21]
30. The Semantic Network: National Library of Medicine - NIH. The UMLS Semantic Network URL: https://semanticnetwork.nlm.nih.gov/ [accessed 2019-08-21]
31. NCBI. 2019. Metathesaurus: Original Release Format (ORF) URL: https://www.ncbi.nlm.nih.gov/books/NBK9682/ [accessed 2019-08-21]
32. Raghavan P, Fosler-Lussier E, Lai A. Learning to Temporally Order Medical Events in Clinical Text. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014 Presented at: ACL'14; July 8, 2012; Jeju Island, Korea p. 70-74. [doi: 10.3115/v1/p14-1094]
33. Tatarinov I, Viglas S, Beyer KS, Shanmugasundaram J, Shekita EJ, Zhang C. Storing and Querying Ordered XML Using a Relational Database System. In: Proceedings of the 2002 ACM SIGMOD international conference on Management of data. 2002 Presented at: SIGMOD'02; June 3-6, 2002; Wisconsin, USA. [doi: 10.1145/564691.564715]
34. Osborne JD, Gyawali B, Solorio T. Evaluation of YTEX and MetaMap for clinical concept recognition. arXiv preprint 2014:- epub ahead of print - 1402.1668 [FREE Full text]

Abbreviations

APL: average path length
APL_SYM: symmetric average path length
BoC: bag-of-concept
CA: common ancestor
DAG: directed acyclic graph
EHR: electronic health record
FN: false-negative
FP: false-positive
ICU: intensive care unit
MIMIC: Multiparameter Intelligent Monitoring in Intensive Care
SNOMED-CT: systemized nomenclature of MEDical clinical terms
TN: true-negative
TP: true-positive
UMLS: unified medical language system

Edited by G Eysenbach; submitted 28.08.19; peer-reviewed by A Gupta, F Jain; comments to author 21.10.19; revised version received 01.03.20; accepted 28.03.20; published 17.07.20.

Please cite as:
Le N, Wiley M, Loza A, Hristidis V, El-Kareh R
Prediction of Medical Concepts in Electronic Health Records: Similar Patient Analysis
JMIR Med Inform 2020;8(7):e16008
URL: https://medinform.jmir.org/2020/7/e16008
doi:10.2196/16008
PMID:32706678

Original Paper

Embedding "Smart" Disease Coding Within Routine Electronic Medical Record Workflow: Prospective Single-Arm Trial

Dee Mangin[1], MBChB, DPH, MRNZCGP, FRNZCGP; Jennifer Lawson[1], BSc, MLIS; Krzysztof Adamczyk[1], BSc; Dale Guenter[1], MD, MPH, CCFP, FCFP

Department of Family Medicine, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Dee Mangin, MBChB, DPH, MRNZCGP, FRNZCGP
Department of Family Medicine
McMaster University
David Braley Health Sciences Centre, 5th Floor
100 Main Street West
Hamilton, ON, L8P 1H6
Canada
Phone: 1 905 525 9140 ext 21219
Email: mangind@mcmaster.ca

Abstract

Background: Electronic medical record (EMR) chronic disease measurement can help direct primary care prevention and treatment strategies and plan health services resource management. Incomplete data and poor consistency of coded disease values within EMR problem lists are widespread issues that limit primary and secondary uses of these data. These issues were shared by the McMaster University Sentinel and Information Collaboration (MUSIC), a primary care practice-based research network (PBRN) located in Hamilton, Ontario, Canada.

Objective: We sought to develop and evaluate the effectiveness of new EMR interface tools aimed at improving the quantity and the consistency of disease codes recorded within the disease registry across the MUSIC PBRN.

Methods: We used a single-arm prospective trial design with preintervention and postintervention data analysis to assess the effect of the intervention on disease recording volume and quality. The MUSIC network holds data on over 75,080 patients, 37,212 currently rostered. There were 4 MUSIC network clinician champions involved in gap analysis of the disease coding process and in the iterative design of new interface tools. We leveraged terminology standards and factored EMR workflow and usability into a new interface solution that aimed to optimize code selection volume and quality while minimizing physician time burden. The intervention was integrated as part of usual clinical workflow during routine billing activities.

Results: After implementation of the new interface (June 25, 2017), we assessed the disease registry codes at 3 and 6 months (intervention period) to compare their volume and quality to preintervention levels (baseline period). A total of 17,496 International Classification of Diseases, 9th Revision (ICD9) code values were recorded in the disease registry during the 11.5-year (2006 to mid-2017) baseline period. A large gain in disease recording occurred in the intervention period (8516/17,496, 48.67% over baseline), resulting in a total of 26,774 codes. The coding rate increased by a factor of 11.2, averaging 1419 codes per month over the baseline average rate of 127 codes per month. The proportion of preferred ICD9 codes increased by 17.03% in the intervention period (11,007/17,496, 62.91% vs 7417/9278, 79.94%; χ^2_1=819.4; P<.001). A total of 45.03% (4178/9278) of disease codes were entered by way of the new screen prompt tools, with significant increases between quarters (Jul-Sep: 2507/6140, 40.83% vs Oct-Dec: 1671/3148, 53.08%; χ^2_1=126.2; P<.001).

Conclusions: The introduction of clinician co-designed, workflow-embedded disease coding tools is a very effective solution to the issues of poor disease coding and quality in EMRs. The substantial effectiveness in a routine care environment demonstrates usability, and the intervention detail described here should be generalizable to any setting. Significant improvements in problem list coding within primary care EMRs can be realized with minimal disruption to routine clinical workflow.

(JMIR Med Inform 2020;8(7):e16764) doi:10.2196/16764

KEYWORDS

chronic disease management; comorbidity; problem list; disease coding; disease registry; data improvement; electronic medical record; electronic health record; practice-based research network; population health; primary care; family medicine

Introduction

Primary care is at the center of health care delivery and coordination and is critically positioned to achieve better population health outcomes and address health inequity within clinical care [1,2]. Chronic disease and multimorbidity are increasingly prevalent in primary care populations [3-6]. Chronic disease identification at the individual level helps to inform better patient care and flags the potential burden of illness and of patients' care experience. Chronic disease measurement at the practice and population level can help direct prevention strategies and plan health services resource management [3,4,7,8].

The uptake of electronic medical records (EMRs) internationally is high [9]. In Canada, 83% of primary care physicians are using EMRs [10]. Data within primary care EMRs support care for the individual patient. Aggregated, these data may also support practice-based and population health initiatives to understand, target, and deliver care [11], supporting both epidemiological research and quality improvement [11-13]. The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is one of several national networks that aggregate EMR data to support this work [7,8,14,15]. However, data completeness and consistency of coded values within EMR problem lists or disease registries limit primary and secondary uses of these data [4,16-20].

Primary care clinicians manage, on average, 3 problems per 10- to 15-minute consultation. They have limited time to devote to clinical encounter tasks and even less time for additional data recording and quality tasks that do not relate to individual patient care workflow [21,22]. Primary care physicians spend around half of their clinic time and 1 to 2 hours of after-clinic work devoted to EMR tasks [21,22]. Administrative tasks, including billing, account for around half of the time spent interacting with the EMR.

Primary care practice-based research networks (PBRNs) are clinician collectives focused on asking and answering research questions relevant to their practice context, often using aggregate, routinely collected EMR data. A PBRN offers an ideal setting to imagine and trial interventions that could improve data quality, while not interrupting clinician workflow.

The McMaster University Sentinel and Information Collaboration (MUSIC) PBRN in Hamilton, Ontario, Canada, contributes deidentified EMR data to the CPCSSN national network. Validated algorithms estimate chronic disease prevalence using disease registry codes, billing codes, and medication data [23]. The MUSIC network showed a low prevalence and variability in disease registry codes in relation to the patient population being served.

Our network has been previously successful in implementing an automated, electronic sentinel influenza reporting program integrated into the EMR [24]. We hypothesized that, if co-designed with clinicians, embedding "smart" disease recording within usual EMR clinical workflow could improve disease registry coding volume and quality without any significant burden for clinicians. In this paper, we describe the design, development, and results of a trial of implementation of disease coding tools within the EMR on disease code volume and consistency.

Methods

We conducted a pragmatic trial of an intervention aimed at improving the quantity and the consistency of coded disease data recorded within the disease registry across the MUSIC PBRN.

Setting

The study was set within the MUSIC practice-based research network. The MUSIC network holds data on over 75,080 patients, 37,212 currently rostered, from a broad range of neighborhoods within Hamilton, Ontario, Canada, and the surrounding area. All clinicians use the open source EMR, Open Source Clinical Application and Resources (OSCAR).

Study Design

We used a single-arm prospective trial design with preintervention and postintervention data analysis to assess the effect of the intervention on disease recording volume and quality.

Intervention Development

We discussed the project rationale with project stakeholders, including clinicians, clinic executives, and MUSIC network staff, to establish project support. There were 5 key aspects to our intervention development: literature review, as-is state investigation of the EMR interface, user engagement in design, standardization of disease codes, and iterative prototype feedback cycles.

Literature Review

We first conducted a nonexhaustive literature review to inform the interface design, noting barriers and facilitators for EMR meaningful use [13,25-28]. Prior research demonstrated the concept of leveraging billing workflow for disease-related data improvement [18] and the disease code morbidities most relevant to primary care [27].

As-Is State Investigation

The research team investigated the EMR interface for disease data capture within the OSCAR disease registry and within the billing module. Multiple disease registry issues were flagged, including the poor visibility of disease recording tools, which required side-stepped navigation. International Classification of Diseases, 9th Revision (ICD9) code selection was cumbersome due to nonintuitive term names arranged in a large, flat list that lacked organization.

The billing module is an obligatory part of clinical workflow and requires use of provincially issued diagnostic billing codes. The disease coding component of the billing module was explored for its capacity to be leveraged in disease registry code capture, and challenges to this plan were detected. Similar to the ICD9 coding tools, tools for selecting billing codes lacked clinician-friendly naming, quick-pick lists, or an easy method for search and selection of common conditions. Provincial diagnostic billing codes often lacked specificity, bundling several related conditions together, precluding their use in specific disease identification. Of particular note, the last inputted diagnostic code used to bill the previous patient encounter remained populated in the field, satisfying that portion of the data entry criteria for the billing process and providing little incentive for clinicians to choose the diagnostic code best matched to the current patient encounter.

User Engagement in Design

We engaged 4 clinicians as project advisors and champions. Semistructured interviews with champions identified issues that were possibly contributing to the low volume of disease registry codes and lack of code consistency; these fell into categories of people (physician users), process (workflow and optimized use), and technology (interface).

Stated issues included lack of awareness of how to optimally use disease coding tools, along with time constraints related to clinical workflows and data collection activities. Champions noted a lack of confidence in optimal code selection for both billing codes and disease registry codes, as coding tools were not well supported with search and retrieval tools or quick-pick lists that featured organized and complete sets of preferred terms presented in clinician-friendly formats. Issues of time inefficiency and workflow redundancy related to the need to separately select ICD9 code values for the disease registry when a billing diagnostic code value is already mandated for creating a billing invoice. Champions also reasoned that a firm clinical diagnosis does not always occur at the patient's first billed encounter for the problem. Disease registry interface issues identified by physicians echoed many of the same constraints and barriers that researchers noted during the as-is state investigation, including low visibility of the disease registry module within the EMR and its lack of integration within clinical documentation workflow.

Standardized Disease Codes

We found that the terminology standard, Clinician-Friendly Pick-List Guide for clinical assessment [29], offered for licensed use from the Canadian Institute for Health Information (CIHI), provided a good basis for composing clinician-friendly, chronic disease quick-pick lists for both the billing diagnostic codes and the disease registry codes. We created a reference table composed of 1:1 matches between provincial diagnostic billing codes and the best equivalent ICD9 code to be leveraged for disease registry code capture in the new interface solution (Multimedia Appendix 1).

Iterative Design and Feedback Cycles

We developed wire-framed interface prototypes designed to address clinician-noted EMR interface constraints and to increase integration of the disease registry coding into the routine billing process workflow. We sought prototype feedback from clinical champions on (1) the selection of specific codes and their outward-facing names within quick-pick lists, (2) the interface ease of use and its fit into the clinical documentation workflow, and (3) the comprehensibility of data coding interface inputs, screen prompts, and outputs.

The OSCAR EMR service provider contributed substantially to the development of design features that were mindful of the constraints of the EMR platform. A functioning prototype of the interface solution was hosted on a project server and presented to the larger group of clinician end users, with support by clinical champions. This step allowed for consideration of other important design perspectives that were factored into the final interface solution and training of clinician end users.

Intervention Description

The final EMR interface solution (Figures 1 and 2) addressed the key issues identified by champions, incorporating disease coding prompts within usual workflow, ease of use, and minimal time burden.

Figure 1. The quick-pick list for disease registry data entry with a pop-up prompt embedded within the billing module.

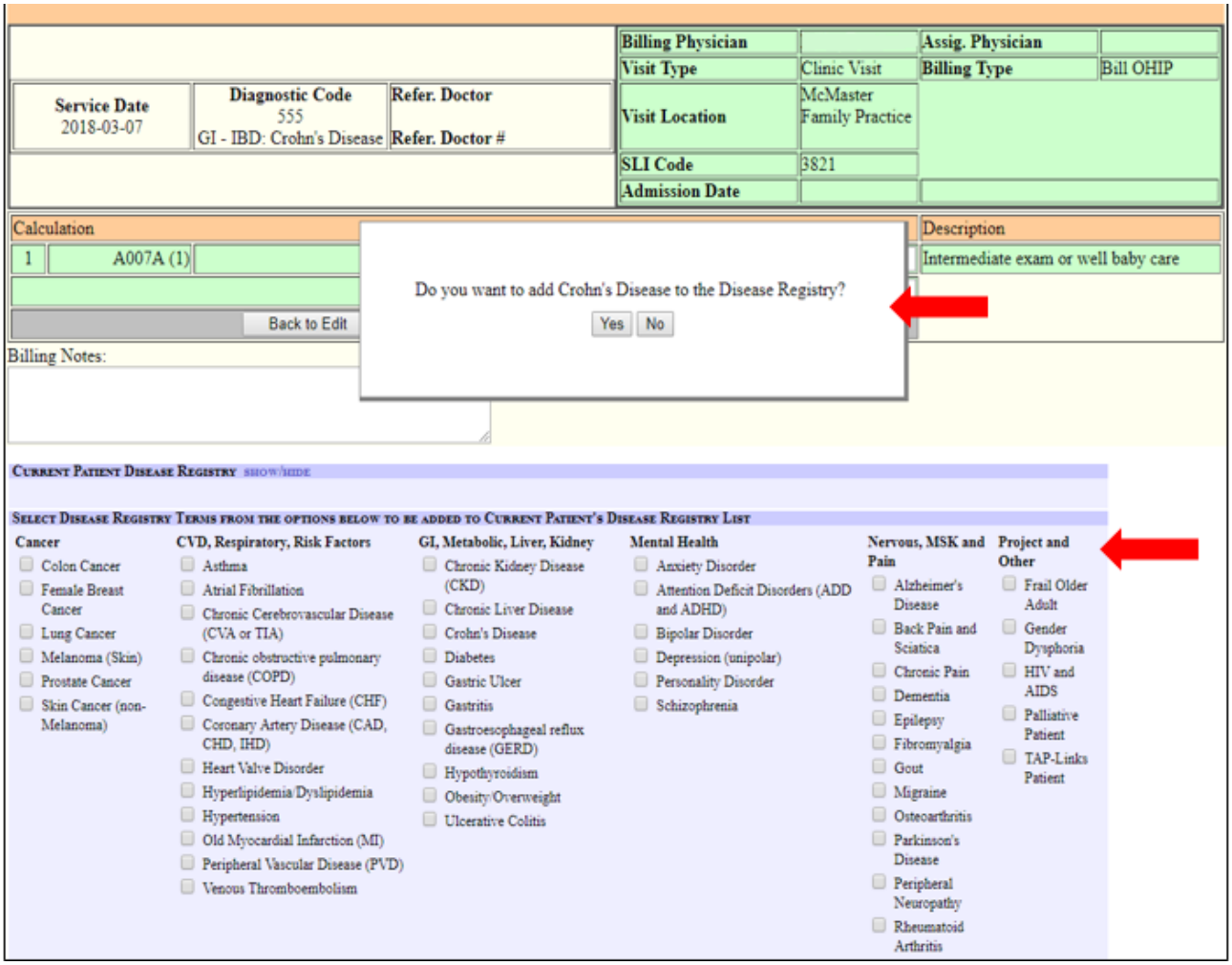


Figure 2. Screenshot of the billing diagnostic quick-pick list.

DxCode	Description
174	Cancer - Breast (Female)
153	Cancer - Colon, Large Intestine, Not rectum
162	Cancer - Lung, Bronchus
172	Cancer - Melanoma (Skin)
173	Cancer - Non-melanoma (Skin), BCC, SCC
185	Cancer - Prostate
585	CKD, Chronic Kidney Disease
530	GI - GERD, Esophageal Disorders
555	GI - IBD: Crohn's Disease
556	GI - IBD: Ulcerative Colitis
571	GI - Liver Disease (Chronic), Failure, Cirrhosis, Fatty
531	GI - Peptic Ulcer, Gastric Ulcer
427	Heart - A Fib, Arrythmia
428	Heart - CHF, Congestive Heart Failure
394	Heart Valve Disorder, Valve Stenosis, Insufficiency
042	HIV and AIDS
314	Mental - ADD, ADHD, Attention Deficit Disorder
300	Mental - Anxiety Disorders, OCD
296	Mental - Bipolar Disorder
311	Mental - Depression (unipolar)
301	Mental - Personality Disorder
295	Mental - Schizophrenia
309	Mental - Stressed, Adjustment Reaction
250	Metab - Diabetes or Diabetes Complications
244	Metab - Hypothyroid (Acquired)
278	Metab - Obesity, Overweight
726	MSK - Fibromyalgia
274	MSK - Gout
724	MSK - Low Back Pain, Sciatica
715	MSK - OA, Osteoarthritis
714	MSK - RA, Rheumatoid Arthritis, Still's disease
331	Neuro - Dementia, Alzheimer's, Lewy Body
345	Neuro - Epilepsy, Seizure
346	Neuro - Migraine
332	Neuro - Parkinson's Disease
356	Neuro - Peripheral Neuralgia, Neuritis, Carpal Tunnel
290	Neuro - Vascular Dementia, Cognitive, Memory Impairment
493	Resp - Asthma, Reactive Airways, Allergic Bronchitis
496	Resp - COPD, Chronic Obstructive Pulmonary Disease
401	Risk - HTN, Hypertension
272	Risk - Hyperlipidemia, Dyslipidemia
413	Vascular - Angina Pectoris, Acute CAD, Acute MI
412	Vascular - Chronic CAD, Arteriosclerosis, IHD, CHD, Old
437	Vascular - Chronic CVA, Chronic Cerebrovascular Disease
443	Vascular - PVD, Claudication, Raynaud's Disease

Disease Code Quick-Pick Lists

We renamed the ICD disease registry codes with 51 front-facing clinician-friendly terms for common chronic conditions in primary care, guided by the CIHI list and clinical champion feedback. We organized the codes into a quick-pick list with clinically logical groupings and inserted this within the billing module (Figure 1) and the disease registry module. A total of 44 billing diagnostic codes were selected for closest equivalence to the disease registry codes (Multimedia Appendix 1) and fitted with new clinician-friendly term names. Where codes comprised multiple conditions, the one most relevant to the matched ICD9 code formed the leading portion of the term name. These were presented as an easily accessible drop-down list (quick-pick list) within the billing module to be used during obligatory billing activities (Figure 2).

Disease Registry Code Prompt Within the Billing Module

The table of billing diagnostic codes matched to ICD9 disease registry codes was posted to the back end of the EMR for automatic nomination of an equivalent disease registry ICD9 code via a pop-up window prompt (Figure 1). The timing of the prompt coincides with clinical cognitive processes around diagnosis and obligatory billing documentation tasks for clinical encounters. When one of the quick-pick billing diagnostic codes is selected, a pop-up screen appears that asks, “Do you want to add [term name] to the disease registry?” with “Yes” and “No” button selections. If the matching ICD9 code value is already in the patient’s disease registry, no prompt is presented. Clicking on “Yes” adds the underlying ICD9 code value to the patient’s disease registry. If “No” is clicked and the same billing code for the same patient is selected at a later consultation, the screen prompt is presented again up to 3 times, after which it is no longer presented. This repeated prompt was suggested by the clinician advisors who gave feedback that diagnosis is not

always confirmed at the first presentation for a condition and that 3 times offers a reasonable opportunity to select a disease code without creating undue burden or contributing to alert fatigue.

Once the billing module interface changes were implemented, each clinic site hosted group training sessions for clinician end users that reinforced project rationale and described optimized use of new interface features. Clinician champions at each site encouraged and supported their peers in using the new tools. End users provided interface experience feedback to the project team via clinician champions.

Outcome Measures

Primary Outcome

The primary outcome was the change in total number of disease registry codes in the MUSIC data set compared with the expected number estimated from the preintervention period to assess whether the intervention had been successful.

Secondary Outcomes

The secondary outcomes were (1) data consistency, assessed by comparing the proportion of ICD9 codes that matched to the preferred codes at baseline and during the 6-month postintervention phase; (2) usability of the new interface coding tools, assessed by comparing counts of the mode by which the new codes were being added (interface prompts versus other means, eg, direct keying in); and (3) patient characteristics, including the number of patients with disease registry codes identified in their records and whether new codes were added to patients' partially completed disease registries or de novo, to patients' disease registries with no previous disease code entries.

Data Collection Period

We implemented the EMR interface changes on June 25, 2017. The preintervention data set includes all disease registry codes added between January 23, 2006, and June 24, 2017 (baseline period). The intervention data set includes all codes collected on or after the implementation date of June 25, 2017 (intervention period).

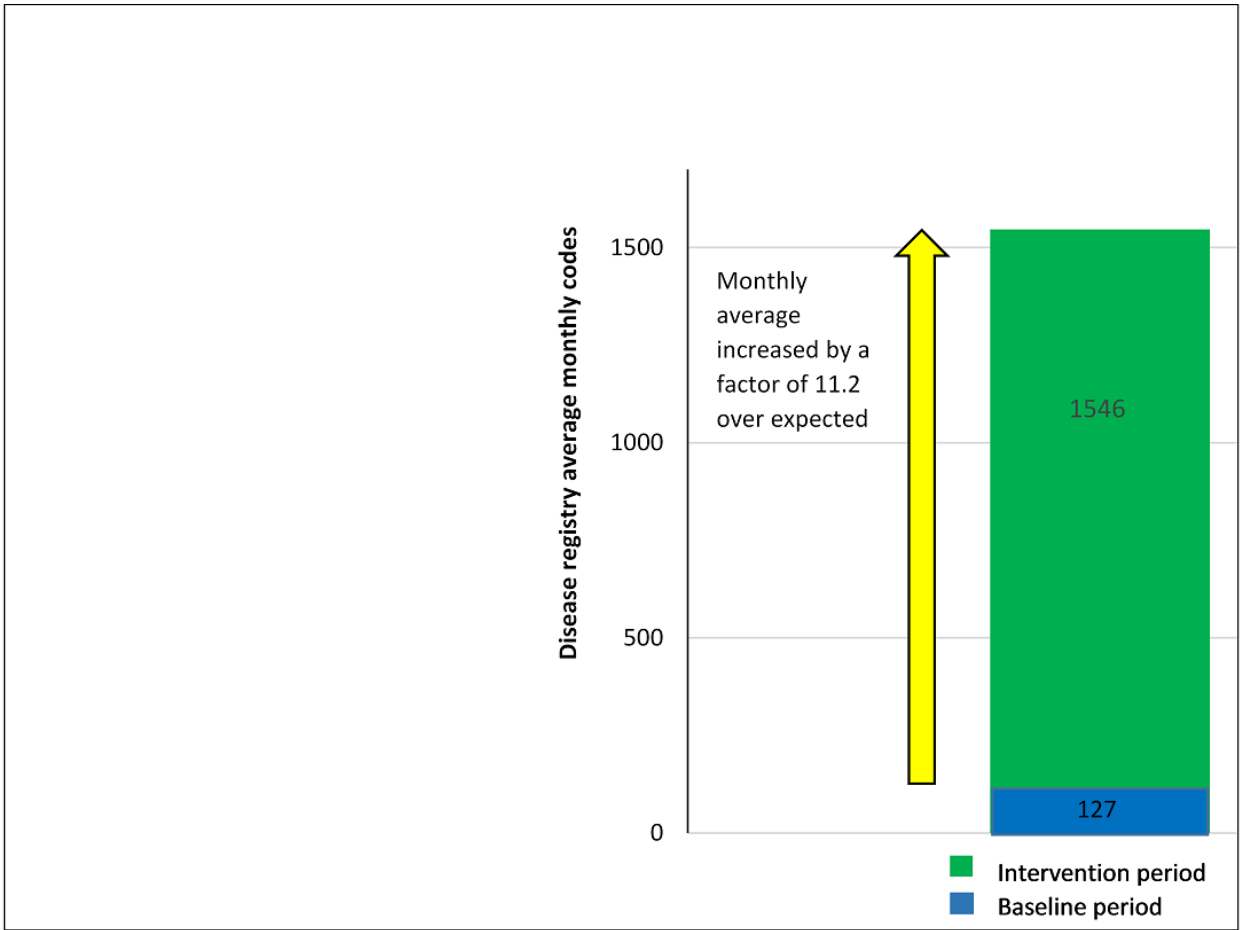
We compared the baseline period codes to the intervention period codes at 3 and 6 months after initiation of the intervention to assess their volume and quality.

Results

Primary Outcome

During the 11-year baseline period (2006 to mid-2017), 17,496 ICD9 code values were recorded in the disease registry. This represents an average code collection rate of 127 codes per month. After implementation of new interface features, 9278 codes were added over 6 months, representing 8516 more codes over the expected volume of 762 codes. Disease registry codes were therefore increased by 48.67% (8516/17,496) by the intervention. The intervention period coding rate averaged 1546 codes per month, which is an increase of 1419 codes per month over the baseline rate (127 codes per month), or a factor of 11.2 (Figure 3). There were more codes added in the first 3 months (6138/9278) of the intervention period compared with the last 3 months (3140/9278).

Figure 3. Disease registry monthly code collection rates of baseline and intervention periods.



	Jan 2006 to Dec 2008	Jan 2009 to Dec 2011	Jan 2012 to Dec 2014	Jan 2015 to Jun 2017	Jul 2017 to Dec 2017
Codes volume, n (%)	5313 (19.84)	2409 (9.00)	3089 (11.54)	6685 (24.97)	9278 (34.65)
Total codes, n (%)	17,496 (65.35)				9278 (34.65)
Expected codes, n	N/A				762
Total additional intervention codes (Observed – Expected), n	N/A				8516
	Proportional increase in codes over baseline				**48.67%**

NA: Not applicable

Secondary Outcomes

Data Consistency

We found a statistically significant percentage point increase of 17.03% (χ^2_1=819.4; *P*<.001) in the proportion of preferred ICD9 codes selected in the intervention period (7417/9278, 79.94%) compared with the baseline period (11,007/17,496, 62.91%) (Table 1). This shifted the proportion of preferred ICD codes overall from 62.91% (11,007/17,496) to 68.81% (18,424/26,774).

Table 1. Proportion of preferred International Classification of Diseases, 9th Revision codes used in the baseline and intervention periods.

Period	Preferred ICD[a] term codes, n (%) (n=18,424)[b]	Nonpreferred ICD term codes, n (%) (n=8350)[c]	Total codes, n (N=26,774)
Baseline period	11,007 (62.91)	6489 (37.09)	17,496
Postintervention period	7417 (79.94)	1861 (20.06)	9278
Proportional change	3590 (17.03)	4628 (–17.03)	N/A[d]

[a]ICD: International Classification of Diseases.

[b]68.81% of total codes.

[c]31.19% of total codes.

[d]N/A: not applicable.

Usability of Coding Tools

Over the 6-month follow-up period, 45.03% (4178/9278) of codes were added via the new screen prompt triggered by the quick-pick list billing codes, with a significant rise in proportion from the first 3 months to the last 3 months (2507/6140, 40.83% vs 1671/3148, 53.08%; χ^2=126.2; *P*<.001). The remaining codes were directly added through (1) the quick-pick list of 51 clinician-friendly disease registry terms positioned within the final screen of the billing module, (2) the quick-pick list in the disease registry module itself, or (3) manually typing the selected codes into the designated field of the disease registry module.

Patient Characteristics

A total of 12,459 unique patients had one or more disease registry codes in their record; 28.78% (3486/12,459) had codes recorded during the postintervention period. Among these 3486 patients with postintervention codes, 1527 (43.80%) had no previous disease registry codes in their record, indicating that the new disease coding tools were balanced between extending partially completed disease registries and creating new registries for patients (Multimedia Appendix 1). Demographic characteristics of patients with disease registry coding can be found in Multimedia Appendix 1.

Discussion

Principal Results

Our study demonstrates that embedding clinician co-designed EMR disease recording tools into routine workflow, reinforced by training and peer support, results in substantial improvements in the quantity and quality of disease registry coding. In just 6 months, we found an absolute increase of 53.03% (9278/17,496), or a 48.67% (8516/17,496) gain over the number of disease codes expected from the previous 11-year period. There were more codes added in the first 3 months of the intervention period compared with the second 3 months. We saw an increase in the second 3 months in the proportion of codes being added via the new screen prompt triggered by the billing diagnosis code for that encounter. These findings might be expected; the potential gap in disease registry coding narrows as codes are added to a given patient's problem list for existing but uncoded diseases, so eventually only new disorders identified at subsequent encounters need to be added.

The consistency of codes also increased, with a greater selection of preferred codes added to the disease registry within the intervention period compared with the baseline period. Having a more consistent set of disease codes improves the quality and thereby the value of the data set, supporting both population health research and quality improvement initiatives. The use of the new tools over the older, less systematic ways of entering disease registry codes suggests that this is an acceptable way to substantially increase disease coding and quality.

Strengths

We used a pragmatic, iterative approach to a primary care EMR enhancement project, with clinician end users involved in design at each step. We applied multiple methods to thoroughly inform the design, including potential solutions from the literature, a national reference standard, and the local EMR service provider. The solution was fitted to routine clinical documentation workflow to limit burden on clinicians. The 6-month follow-up provides a useful and informative assessment of the longitudinal benefit of the intervention. With the pace of change in health informatics, in addition to shifts in definitions for billing codes, gathering follow-up data over this targeted period avoids most potential process and contextual confounders.

Limitations

While the 6-month evaluation period avoids the confounders highlighted above, it also provides a limited scope with which to measure the long-term success of the interface change. Further longitudinal evaluation will help illuminate any extinction of effect as the coding gap closes and whether the predicted further increase in the overall consistency of codes is supported by the data.

This solution of prompting physicians to add disease registry codes as part of the billing documentation workflow limits coding to patients attending medical appointments. Other solutions for completing the disease registry for patients who attend infrequently will need to be devised to ensure representative problem list data for this group. Disease registry back coding of patients using validated algorithmic case definitions (eg, those offered by CPCSSN [23]) integrated with clinician input may offer a further opportunity to assign missing disease registry codes to inactive patients.

The intervention development and implementation had 5 key aspects of design, as well as training and peer support in

implementation. It is not possible to determine the relative contribution of each to the overall effectiveness.

Comparison With Prior Work

Leading electronic health researchers have identified knowledge and research gaps in primary care EMRs, specifically the need for reliable disease and multimorbidity metrics to inform optimal management of patients' clinical problems and population-level health strategies [30]. These issues were addressed in this research, first with identification of EMR design constraints affecting disease coding, followed by development, implementation, and evaluation of new data collection tools toward improved data quantity and quality.

Similar to other reported findings [17,19,31,32], we identified data quality issues in the MUSIC EMR data set that limit confidence in the use of chronic disease data for practice-based initiatives and research. Previous research in problem list design identified the benefit of incorporating the problem list into the clinical documentation routine [18,26]; this need was echoed in the feedback from MUSIC clinicians that were consulted in the design of the EMR interface improvement.

EMR usability studies have generated a myriad of clinician observations that identify navigation, safety, and cognitive load issues associated with EMRs [33]. This research underscores the importance of clinician input in EMR design and redesign projects. Continuous engagement of clinician end users in EMR implementation projects [34] or EMR use enhancement projects [35,36] has previously been reported to increase the projects' likelihood of success [37]. Clinicians in the role of project champions and change management agents have proven essential for the encouragement of advanced EMR feature use [38].

In our study, the application of local physician co-design, which saw key clinician input into solution development, implementation planning, training components, and championing of new coding features, conceivably translated into an interface solution reasonably fitted to clinician workflow, leading to acceptability and uptake. Our study demonstrates that development and delivery of a relevant and usable solution for improving chronic disease recording is attainable.

Conclusion

Our pragmatic approach to EMR interface redesign resulted in substantial gains in disease code quantity and quality, providing a much-improved data set for asking and answering clinically important research questions. Clinician involvement in the intervention design, training, and peer support resulted in an accepted solution that placed little burden on clinicians. The often used quote, "If we want evidence-based practice, we need practice-based evidence" [39] mandates that PBRN data quality and quantity are adequate for this task. The study demonstrates that achieving significant improvements in problem list coding within primary care EMRs can be realized with minimal disruption to routine clinical workflow.

Acknowledgments

The authors would like to thank the primary care clinicians and patients of the MUSIC PBRN who contribute their data to the network through which the study data were generated and willingly contributed to this project. We also acknowledge Krzysztof Adamczyk, the information technology lead for the MUSIC network, and Ronnie Cheng, OSCAR program developer for the MUSIC network. We thank Kathy De Caire, Kati Ivanyi, Doug Oliver, and Jill Berridge for their executive support, and Casey Irvin for his help in the creation of figures and tables. We acknowledge the support of the McMaster University Department of Family Medicine for this PBRN.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary Tables 1-5.
[DOCX File , 17 KB - medinform_v8i7e16764_app1.docx]

References

1. Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. Milbank Q 2005;83(3):457-502 [FREE Full text] [doi: 10.1111/j.1468-0009.2005.00409.x] [Medline: 16202000]
2. Starfield B, Gérvas J, Mangin D. Clinical care and health disparities. Annu Rev Public Health 2012 Apr;33:89-106. [doi: 10.1146/annurev-publhealth-031811-124528] [Medline: 22224892]
3. Smith SM, Wallace E, O'Dowd T, Fortin M. Interventions for improving outcomes in patients with multimorbidity in primary care and community settings. Cochrane Database Syst Rev 2016 Mar 14;3:CD006560 [FREE Full text] [doi: 10.1002/14651858.CD006560.pub3] [Medline: 26976529]
4. Chronic Disease Management in Primary Health Care: A Demonstration of EMR Data for Quality and Health System Monitoring. Canadian Institute for Health Information. 2014 Jan. URL: https://secure.cihi.ca/free_products/Burden-of-Chronic-Diseases_PHC_2014_AiB_EN-web.pdf [accessed 2019-10-09]

5. Seniors and the Health Care System: What Is the Impact of Multiple Chronic Conditions? Canadian Institute for Health Information. 2011 Jan. URL: https://secure.cihi.ca/free_products/air-chronic_disease_aib_en.pdf [accessed 2020-07-07] [WebCite Cache ID 6PBBB6Bk0]
6. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet 2012 Jul 7;380(9836):37-43 [FREE Full text] [doi: 10.1016/S0140-6736(12)60240-2] [Medline: 22579043]
7. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PLoS One 2014;9(6):e99825 [FREE Full text] [doi: 10.1371/journal.pone.0099825] [Medline: 24941260]
8. Nicholson K, Terry AL, Fortin M, Williamson T, Thind A. Understanding multimorbidity in primary health care. Can Fam Physician 2015 Oct;61(10):918, e489-918, e490 [FREE Full text] [Medline: 26472799]
9. Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff (Millwood) 2012 Dec;31(12):2805-2816 [FREE Full text] [doi: 10.1377/hlthaff.2012.0884] [Medline: 23154997]
10. 2018 Canadian Physician Survey. Canada Health Infoway. 2018 Dec. URL: https://infoway-inforoute.ca/en/component/edocman/resources/reports/benefits-evaluation/3643-2018-canadian-physician-survey [accessed 2020-07-07]
11. Vaghefi I, Hughes JB, Law S, Lortie M, Leaver C, Lapointe L. Understanding the Impact of Electronic Medical Record Use on Practice-Based Population Health Management: A Mixed-Method Study. JMIR Med Inform 2016 Apr 04;4(2):e10 [FREE Full text] [doi: 10.2196/medinform.4577] [Medline: 27044411]
12. Gentil M, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, et al. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. BMC Med Inform Decis Mak 2017 Sep 25;17(1):139 [FREE Full text] [doi: 10.1186/s12911-017-0538-x] [Medline: 28946908]
13. Paré G, Raymond L, Guinea AOD, Poba-Nzaou P, Trudel M, Marsan J, et al. Electronic health record usage behaviors in primary care medical practices: A survey of family physicians in Canada. Int J Med Inform 2015 Oct;84(10):857-867. [doi: 10.1016/j.ijmedinf.2015.07.005] [Medline: 26238705]
14. Birtwhistle R, Queenan JA. Update from CPCSSN. Can Fam Physician 2016 Oct;62(10):851 [FREE Full text] [Medline: 27737985]
15. Birtwhistle R, Keshavjee K, Lambert-Lanning A, Godwin M, Greiver M, Manca D, et al. Building a pan-Canadian primary care sentinel surveillance network: initial development and moving forward. J Am Board Fam Med 2009;22(4):412-422 [FREE Full text] [doi: 10.3122/jabfm.2009.04.090081] [Medline: 19587256]
16. Coleman N, Halas G, Peeler W, Casaclang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. BMC Fam Pract 2015 Feb 05;16:11 [FREE Full text] [doi: 10.1186/s12875-015-0223-z] [Medline: 25649201]
17. Singer A, Yakubovich S, Kroeker AL, Dufault B, Duarte R, Katz A. Data quality of electronic medical records in Manitoba: do problem lists accurately reflect chronic disease billing diagnoses? J Am Med Inform Assoc 2016 Nov;23(6):1107-1112. [doi: 10.1093/jamia/ocw013] [Medline: 27107454]
18. Wright A, McCoy AB, Hickman TT, Hilaire DS, Borbolla D, Bowes WA, et al. Problem list completeness in electronic health records: A multi-site study and assessment of success factors. Int J Med Inform 2015 Oct;84(10):784-790 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.06.011] [Medline: 26228650]
19. Greiver M, Sullivan F, Kalia S, Aliarzadeh B, Sharma D, Bernard S, et al. Agreement between hospital and primary care on diagnostic labeling for COPD and heart failure in Toronto, Canada: a cross-sectional observational study. NPJ Prim Care Respir Med 2018 Mar 09;28(1):9 [FREE Full text] [doi: 10.1038/s41533-018-0076-8] [Medline: 29523779]
20. Greiver M, Wintemute K, Aliarzadeh B, Martin K, Khan S, Jackson D, et al. Implementation of data management and effect on chronic disease coding in a primary care organisation: A parallel cohort observational study. J Innov Health Inform 2016 Oct 12;23(3):843 [FREE Full text] [doi: 10.14236/jhi.v23i3.843] [Medline: 28059692]
21. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med 2016 Dec 06;165(11):753-760. [doi: 10.7326/M16-0961] [Medline: 27595430]
22. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W, Sinsky CA, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
23. Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. Ann Fam Med 2014 Jul;12(4):367-372 [FREE Full text] [doi: 10.1370/afm.1644] [Medline: 25024246]
24. Price D, Chan D, Greaves N. Physician surveillance of influenza: collaboration between primary care and public health. Can Fam Physician 2014 Jan;60(1):e7-15 [FREE Full text] [Medline: 24452584]
25. Chowdhry SM, Mishuris RG, Mann D. Problem-oriented charting: A review. Int J Med Inform 2017 Jul;103:95-102. [doi: 10.1016/j.ijmedinf.2017.04.016] [Medline: 28551008]

26. Simons SMJ, Cillessen FHJM, Hazelzet JA. Determinants of a successful problem list to support the implementation of the problem-oriented medical record according to recent literature. BMC Med Inform Decis Mak 2016 Aug 02;16:102 [FREE Full text] [doi: 10.1186/s12911-016-0341-0] [Medline: 27485127]
27. Tonelli M, Wiebe N, Fortin M, Guthrie B, Hemmelgarn BR, James MT, Alberta Kidney Disease Network. Methods for identifying 30 chronic conditions: application to administrative data. BMC Med Inform Decis Mak 2015 Apr 17;15:31 [FREE Full text] [doi: 10.1186/s12911-015-0155-5] [Medline: 25886580]
28. Rahal RM, Mercer J, Kuziemsky C, Yaya S. Primary Care Physicians' Experience Using Advanced Electronic Medical Record Features to Support Chronic Disease Prevention and Management: Qualitative Study. JMIR Med Inform 2019 Nov 29;7(4):e13318 [FREE Full text] [doi: 10.2196/13318] [Medline: 31782742]
29. Canadian Institute for Health Information. Clinician-Friendly Pick-List Guide. Pan-Canadian Primary Health Care Electronic Medical Record Content Standard, Version 3. 2014. URL: https://secure.cihi.ca/free_products/PHC_EMR_Content_Standard_V3_PickListGuide_EN.pdf [accessed 2020-07-07]
30. Terry AL, Stewart M, Fortin M, Wong ST, Grava-Gubins I, Ashley L, et al. Stepping Up to the Plate: An Agenda for Research and Policy Action on Electronic Medical Records in Canadian Primary Healthcare. Healthc Policy 2016 Nov;12(2):19-32 [FREE Full text] [Medline: 28032822]
31. Price M, Davies I, Rusk R, Lesperance M, Weber J. Applying STOPP Guidelines in Primary Care Through Electronic Medical Record Decision Support: Randomized Control Trial Highlighting the Importance of Data Quality. JMIR Med Inform 2017 Jun 15;5(2):e15 [FREE Full text] [doi: 10.2196/medinform.6226] [Medline: 28619704]
32. Sollie A, Sijmons RH, Helsper C, Numans ME. Reusability of coded data in the primary care electronic medical record: A dynamic cohort study concerning cancer diagnoses. Int J Med Inform 2017 Mar;99:45-52. [doi: 10.1016/j.ijmedinf.2016.08.004] [Medline: 28118921]
33. Zahabi M, Kaber DB, Swangnetr M. Usability and Safety in Electronic Medical Records Interface Design: A Review of Recent Literature and Guideline Formulation. Hum Factors 2015 Aug;57(5):805-834. [doi: 10.1177/0018720815576827] [Medline: 25850118]
34. Goodison R, Borycki EM, Kushniruk AW. Use of Agile Project Methodology in Health Care IT Implementations: A Scoping Review. Stud Health Technol Inform 2019;257:140-145. [Medline: 30741186]
35. Jones M, Talebi R, Littlejohn J, Bosnic O, Aprile J. An Optimization Program to Help Practices Assess Data Quality and Workflow With Their Electronic Medical Records: Observational Study. JMIR Hum Factors 2018 Dec 21;5(4):e30 [FREE Full text] [doi: 10.2196/humanfactors.9889] [Medline: 30578203]
36. Tran K, Leblanc K, Valentinis A, Kavanagh D, Zahr N, Ivers NM. Evaluating the Usability and Perceived Impact of an Electronic Medical Record Toolkit for Atrial Fibrillation Management in Primary Care: A Mixed-Methods Study Incorporating Human Factors Design. JMIR Hum Factors 2016 Feb 17;3(1):e7 [FREE Full text] [doi: 10.2196/humanfactors.4289] [Medline: 27026394]
37. Gill R, Borycki EM. The Use of Case Studies in Systems Implementations Within Health Care Settings: A Scoping Review. Stud Health Technol Inform 2017;234:142-149. [Medline: 28186031]
38. Terry AL, Ryan BL, McKay S, Oates M, Strong J, McRobert K, et al. Towards optimal electronic medical record use: perspectives of advanced users. Fam Pract 2018 Sep 18;35(5):607-611. [doi: 10.1093/fampra/cmy002] [Medline: 29444228]
39. Green LW. Making research relevant: if it is an evidence-based practice, where's the practice-based evidence? Fam Pract 2008 Dec;25 Suppl 1:i20-i24. [doi: 10.1093/fampra/cmn055] [Medline: 18794201]

Abbreviations

CIHI: Canadian Institute for Health Information
CPCSSN: Canadian Primary Care Sentinel Surveillance Network
EMR: electronic medical record
ICD9: International Classification of Diseases, 9th Revision
MUSIC: McMaster University Sentinel and Information Collaboration
OSCAR: Open Source Clinical Application and Resources
PBRN: practice-based research network

Edited by G Eysenbach; submitted 22.10.19; peer-reviewed by D Gunasekeran, C Fincham; comments to author 17.12.19; revised version received 21.02.20; accepted 10.04.20; published 27.07.20.

Please cite as:
Mangin D, Lawson J, Adamczyk K, Guenter D
Embedding "Smart" Disease Coding Within Routine Electronic Medical Record Workflow: Prospective Single-Arm Trial
JMIR Med Inform 2020;8(7):e16764
URL: http://medinform.jmir.org/2020/7/e16764/
doi:10.2196/16764
PMID:32716304

Original Paper

A Retrospective Analysis of Provider-to-Patient Secure Messages: How Much Are They Increasing, Who Is Doing the Work, and Is the Work Happening After Hours?

Frederick North[1], MD; Kristine E Luhman[2], RN, MBA; Eric A Mallmann[3], BA; Toby J Mallmann[3], BSN; Sidna M Tulledge-Scheitel[1], MD, MPH; Emily J North[4], MD; Jennifer L Pecina[5], MD

[1]Division of Community Internal Medicine, Department of Medicine, Mayo Clinic, Rochester, MN, United States

[2]Mayo Clinic, Rochester, MN, United States

[3]Undergraduate Research Education Program, Mayo Clinic, Rochester, MN, United States

[4]Department of Medicine, NYU Grossman School of Medicine, New York, NY, United States

[5]Department of Family Medicine, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Frederick North, MD
Division of Community Internal Medicine
Department of Medicine
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 5072842511
Email: north.frederick@mayo.edu

Abstract

Background: Patient portal registration and the use of secure messaging are increasing. However, little is known about how the work of responding to and initiating patient messages is distributed among care team members and how these messages may affect work after hours.

Objective: This study aimed to examine the growth of secure messages and determine how the work of provider responses to patient-initiated secure messages and provider-initiated secure messages is distributed across care teams and across work and after-work hours.

Methods: We collected secure messages sent from providers from January 1, 2013, to March 15, 2018, at Mayo Clinic, Rochester, Minnesota, both in response to patient secure messages and provider-initiated secure messages. We examined counts of messages over time, how the work of responding to messages and initiating messages was distributed among health care workers, messages sent per provider, messages per unique patient, and when the work was completed (proportion of messages sent after standard work hours).

Results: Portal registration for patients having clinic visits increased from 33% to 62%, and increasingly more patients and providers were engaged in messaging. Provider message responses to individual patients increased significantly in both primary care and specialty practices. Message responses per specialty physician provider increased from 15 responses per provider per year to 53 responses per provider per year from 2013 to 2018, resulting in a 253% increase. Primary care physician message responses increased from 153 per provider per year to 322 from 2013 to 2018, resulting in a 110% increase. Physicians, nurse practitioners, physician assistants, and registered nurses, all contributed to the substantial increases in the number of messages sent.

Conclusions: Provider-sent secure messages at a large health care institution have increased substantially since implementation of secure messaging between patients and providers. The effort of responding to and initiating messages to patients was distributed across multiple provider categories. The percentage of message responses occurring after hours showed little substantial change over time compared with the overall increase in message volume.

(JMIR Med Inform 2020;8(7):e16521) doi:10.2196/16521

KEYWORDS

patient messages; secure messages; patient portal; provider messages; electronic health records; electronic mail; communication; patients; physicians; physician assistants; nurse practitioners; nurses

Introduction

Background

The volume of secure messages in health care institutions is increasing. A major national survey in 2013 found that 29.6% of the US population had used the internet or email to communicate with a physician or a physician's office in the previous 12 months [1]. More recently, Lee et al [2] found that 37% of customers of a pharmacy chain reported contacting their physicians by email in the last 6 months.

Several health care systems in the United States have examined the increase in the number of messages from patients. Crotty et al [3] noted a tripling in messaging from 2001 to 2010. Cronin et al, Masterman et al, and Shenson et al [4-6] showed dramatic increases in message volumes across multiple specialties, including surgical and pediatric specialties.

Providers have mixed feelings about secure messages. In a survey of 43 clinicians across 5 clinics, 63% disagreed with the statement, "secure messaging reduces my workload," and 33% agreed that "secure messaging has a negative effect on my workflow" [7]. However, 61% agreed that "secure messaging has a positive effect on patient-clinician communication" [7].

Objectives

With providers responsible for an increasing number of secure messages, we looked at how secure messages to patients are distributed among staff at a large health care institution. In addition, with the increasing workload of secure messages, we examined whether there were potential *work after work* issues of using time after normal work hours to complete message responses [8].

Methods

Setting

The study took place at Mayo Clinic, Rochester, Minnesota, United States. Mayo Clinic is a multispecialty clinic with more than 1 million visits annually. There are more than 2200 physicians and scientists at the Rochester, Minnesota, campus.

Mayo Clinic started using Patient Online Services (POS; a secure patient portal) in 2010 for the primary care practice in Rochester, which serves a population of about 140,000. The Mayo Clinic specialty practice started using POS in 2013. The Mayo Clinic specialty practice serves the local community of Rochester and takes referrals from other practices, both nationally and internationally.

The patient portal (POS) at Mayo Clinic gives patients the ability to view their laboratory results, radiology reports, medical images, office and hospital notes, and specialty consultations. In addition, POS has messaging capability for patients and providers to communicate asynchronously by sending messages through a secure server, which also sends these messages to the electronic health record (EHR). Patients must log in securely to POS to send a message. When providers initiate a message or respond to a patient-initiated message through POS, patients are notified by email. To protect privacy, the email notifying the patient of a provider message states that new information is available on their POS (portal) account. In the notification email, patients are given a link to the Mayo POS, but they still need to securely log in to their personal account to view and send these messages. These asynchronous POS messages between patients and providers are what we call secure messages.

Physicians, nurse practitioners (NPs), physician assistants (PAs), and registered nurses (RNs) can receive and respond to secure messages at Mayo Clinic. In addition, licensed practical nurses (LPN) and secretarial staff can also respond to and send secure messages to patients. The categories used in this paper were physician, NP/PA (combined NPs and PAs), RNs, and other (LPN and secretarial).

Provider Secure Message Data Collection

We collected all secure messages sent from the providers at Mayo Clinic, Rochester, from January 1, 2013, to March 15, 2018. The secure message dataset contained the clinic number of the patient, the date and time of day the provider sent the message, the message text, and the identification code of the provider or provider group who sent the message. In addition, the dataset was dichotomously categorized by whether the provider message was a response to a patient-initiated secure message or a provider-initiated secure message. The provider-response message was defined as a reply to a patient message. Provider-initiated messages were messages to patients created de novo by the provider (or created with help from software, as explained below in Abstraction of Provider Messages for Content). The entire secure message dataset contained only mutually exclusive provider-initiated and provider-response messages.

Patient Demographics

To examine the differences between patients who had provider-response messages during the initial time frame and those several years later, we examined demographics of the patients who initiated messages during the first 6 months of the study (January 1, 2013, to June 30, 2013) and patients who initiated messages from the last 6 months (September 15, 2017, to March 15, 2018).

Abstraction of Provider Messages for Content

From a sampling period of October 2017 to February 2018, we randomized and abstracted 1200 messages, 100 each from 12 categories: 3 types of providers (physician, NP/PA, and RN) split into 2 different practices (primary care and specialty), split further into 2 different message types (response and provider initiated). This gave us the 1200 randomized messages of 100 each in 12 categories (3 provider types multiplied by 2 practice types multiplied by 2 message types).

Across the 12 categories, we further dichotomously coded these as being automated or not. We categorized messages as *automated* if the content was completely software generated, such as reminders for screening mammograms and missed appointment notifications. There were other messages initiated that were not completely automated but did not have a personal message. For example, messages reaching out for specific laboratory or imaging tests that required provider input to order, but the message was not personalized to the point of explaining any details about the purpose of the tests. An example of this would be "Your provider has requested the following testing: fasting blood work in March. Please respond with your availability through Patient Online Services." These message types were also categorized as *automated*. Some messages from patients contained only an update that merely required an acknowledgment such as "thanks for the update." When abstracting the content, we also categorized the message responses as containing only an acknowledgment to account for these. Additional information collected was whether there was reference to having consulted another provider. This was to quantify the frequency at which messages could involve more than one provider. An example of this was a message from an NP/PA who wrote: "I spoke again to our C. diff specialist. He would like you to finish 10 days of vancomycin."

It should be noted that Mayo Clinic has a web-based knowledge system called Ask Mayo Expert, which has text-based content and care process algorithms to help with specific clinical questions. Ask Mayo Expert also gives a list of Mayo experts in specific areas. In this case, the NP/PA may have used Ask Mayo Expert to get the name of the Mayo expert in *Clostridium difficile* enteritis. It was outside the scope of this study to see how often providers were using resources such as Ask Mayo Expert (or other web-based resources) to answer patient questions.

Portal Registration and Unique Face-to-Face Patient Visit Counts

Portal registration information was obtained from the Mayo Clinic connected care data. The number of unique patients seen during face-to-face visits was obtained from the Mayo Clinic administrative data.

Work After Work Measure

From the date and time the secure message was completed, we determined whether the messages were sent during the usual business hours of 8 AM to 5 PM from Monday to Friday US central/daylight saving time.

Statistical Analysis

We used JMP version 13.1 statistical analysis software (SAS) for the descriptive statistics as well as for analysis of variance for the differences between messages per patient by year. We used the Cochran-Armitage trend test to examine the trends in proportions, such as the proportion of messages answered outside of Monday to Friday from 8 AM to 5 PM. JMP version 13.1 was used to randomize the selection of messages for abstraction.

Ethics

We excluded all messages from individuals who had not given research authorization. Mayo Clinic sites in Minnesota ask all patients for their research authorization, which is not specific to any individual study. This study was approved by the Mayo Clinic institutional review board (IRB 17-004807).

Results

Message Distribution by Practice Type

A total of 3,941,618 messages were sent by Mayo Clinic providers between January 1, 2013, and March 15, 2018, associated with 353,177 unique patients. We excluded 6.06% (238,870/3,941,618) of the messages from patients who had not given research authorization. After exclusion of the patients without research authorization there were 326,805 unique patients to whom Mayo Clinic providers sent 3,702,748 messages over the study duration. Provider responses to patient-initiated messages accounted for 48.87% (1,809,614/3,702,748) of the messages; provider-initiated messages accounted for 51.13% (1,893,134/3,702,748). The primary care practice accounted for 28.31% (1,048,216/3,702,748) of the messages, and the specialty practices had 71.69% (2,654,532/3,702,748) of the messages. Primary care providers initiated 18.28% (676,674/3,702,748) of the messages and responded to 10.03% (371,542/3,702,748). Specialists initiated 32.85% (1,216,460/3,702,748) and responded to 38.84% (1,438,072/3,702,748).

Practice Volumes and Portal Registration Over Time

The increase in message counts could not be explained by a large growth in patient visits. In fact, the number of unique patients with face-to-face visits each year remained relatively stable, from 365,943 in 2013 to 388,707 in 2017 resulting in a 6% increase. Portal registration in patients with appointments increased from 33% in 2013 to 62% in 2018.

Patient Demographics

Table 1 shows the demographics of the patients who initiated secure messages and had provider responses. The primary care population started portal messages in 2010, whereas specialty practice started in 2013. At least twice as many patient-initiated messages were sent by female patients in both the primary care and specialty practices in 2013, and this female predominance persisted into 2018. Older age groups, especially those ≥65 years, comprised an increasing proportion of the provider-response messages in 2018 compared with 2013.

Table 1. Demographic comparisons of the patients who initiated messages and had provider responses from 2013 to 2018 by provider type (primary care or specialty).

Demographic	Primary care response messages			Specialty response messages		
	First 6 months (2013; n=14,151), n (%)	Last 6 months (2017-2018; n=53,931), n (%)	*P* value[a]	First 6 months (2013; n=20,109), n (%)	Last 6 months (2017-2018; n=115,725), n (%)	*P* value[a]
Age group (years)						
0-17	443 (3.1)	4094 (7.6)	<.001	926 (4.6)	8095 (7.0)	<.001
18-34	2951 (20.9)	9380 (17.4)	<.001	4555 (22.7)	16,730 (14.5)	<.001
35-49	3769 (26.6)	13,796 (25.6)	.01	5473 (27.2)	23,542 (20.3)	<.001
50-64	4903 (34.6)	15,861 (29.4)	<.001	6191 (30.8)	35,721 (30.9)	.82
65-79	1801 (12.7)	8772 (16.3)	<.001	2592 (12.9)	27,220 (23.5)	<.001
≥80	278 (2.0)	2023 (3.8)	<.001	360 (1.8)	4390 (3.8)	<.001
Sex						
Female	9970 (70.5)	34,011 (63.1)	<.001	13,670 (68.0)	65,798 (56.9)	<.001
Male	4175 (29.5)	19,915 (36.9)	<.001	6427 (32.0)	49,898 (43.1)	<.001
Race						
White	13,182 (93.2)	49,424 (91.6)	<.001	18,574 (92.4)	105,646 (91.3)	<.001
Asian	390 (2.8)	1655 (3.1)	.05	535 (2.7)	2614 (2.3)	<.001
Black	151 (1.1)	709 (1.3)	.02	207 (1.0)	1449 (1.3)	.008
Other	428 (3.0)	2143 (4.0)	<.001	793 (3.9)	6016 (5.2)	<.001
Ethnicity						
Hispanic or Latino	199 (1.4)	975 (1.8)	.001	308 (1.5)	2255 (1.9)	<.001
Not Hispanic or Latino	13,641 (96.4)	51,382 (95.3)	<.001	19,102 (95.0)	108,097 (93.4)	<.001
Unknown/not disclosed	311 (2.2)	1574 (2.9)	<.001	699 (3.5)	5373 (4.6)	<.001
Highest level of education						
Some postcollege graduate studies	2251 (15.9)	11,448 (21.2)	<.001	3526 (17.5)	28,092 (24.3)	<.001
4-year college graduate	3752 (26.5)	9208 (17.1)	<.001	5183 (25.8)	21,866 (18.9)	<.001
Some college or 2-year degree	2918 (20.6)	19,690 (36.5)	<.001	4155 (20.7)	37,619 (32.5)	<.001
High school graduate	1454 (10.3)	6003 (11.1)	.004	2127 (10.6)	14,559 (12.6)	<.001
Some high school (did not graduate)	234 (1.7)	1210 (2.2)	<.001	358 (1.8)	992 (0.9)	<.001
Eighth grade or less	16 (0.1)	164 (0.3)	<.001	36 (0.2)	505 (0.4)	<.001
Unknown	3526 (24.9)	6208 (11.5)	<.001	4724 (23.5)	12,092 (10.4)	<.001
Patient's home address						
Minnesota	13,059 (92.3)	52,663 (97.6)	<.001	14,196 (70.6)	68,538 (59.2)	<.001
Contiguous state	357 (2.5)	651 (1.2)	<.001	2053 (10.2)	20,585 (17.8)	<.001
Other (other US states and international)	735 (5.2)	617 (1.1)	<.001	3860 (19.2)	26,602 (23.0)	<.001

[a]Null hypothesis (H0): within primary care or specialty, first 6 months demographic proportion=last 6 months demographic proportion.

Message Counts Over Time

Figure 1 shows the increase in provider messages from 2013 through 2017. Figure 2 shows the rise in distinct patients who received provider messages over the course of the study. Messages per unique patient increased over the 5-year course of the study (Figure 3), and the increase in both provider response to messages and provider-initiated messages per unique patient was statistically significant (Table 2). Figures 4 and 5 show the number of message responses per provider by year for primary care and specialty care.

Figure 1. Message counts by year.

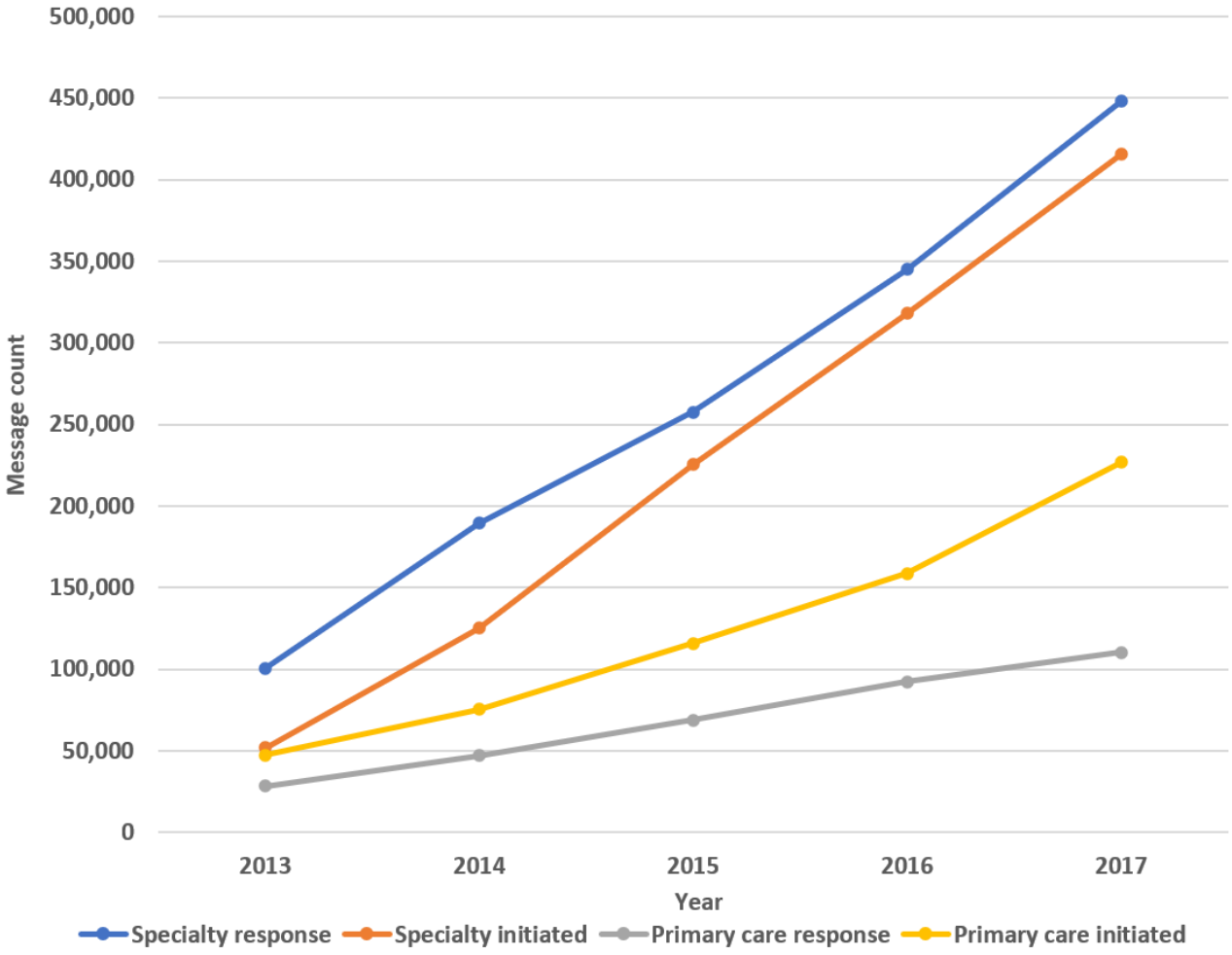


Figure 2. Distinct patients generating a provider response or provider-initiated message by year.

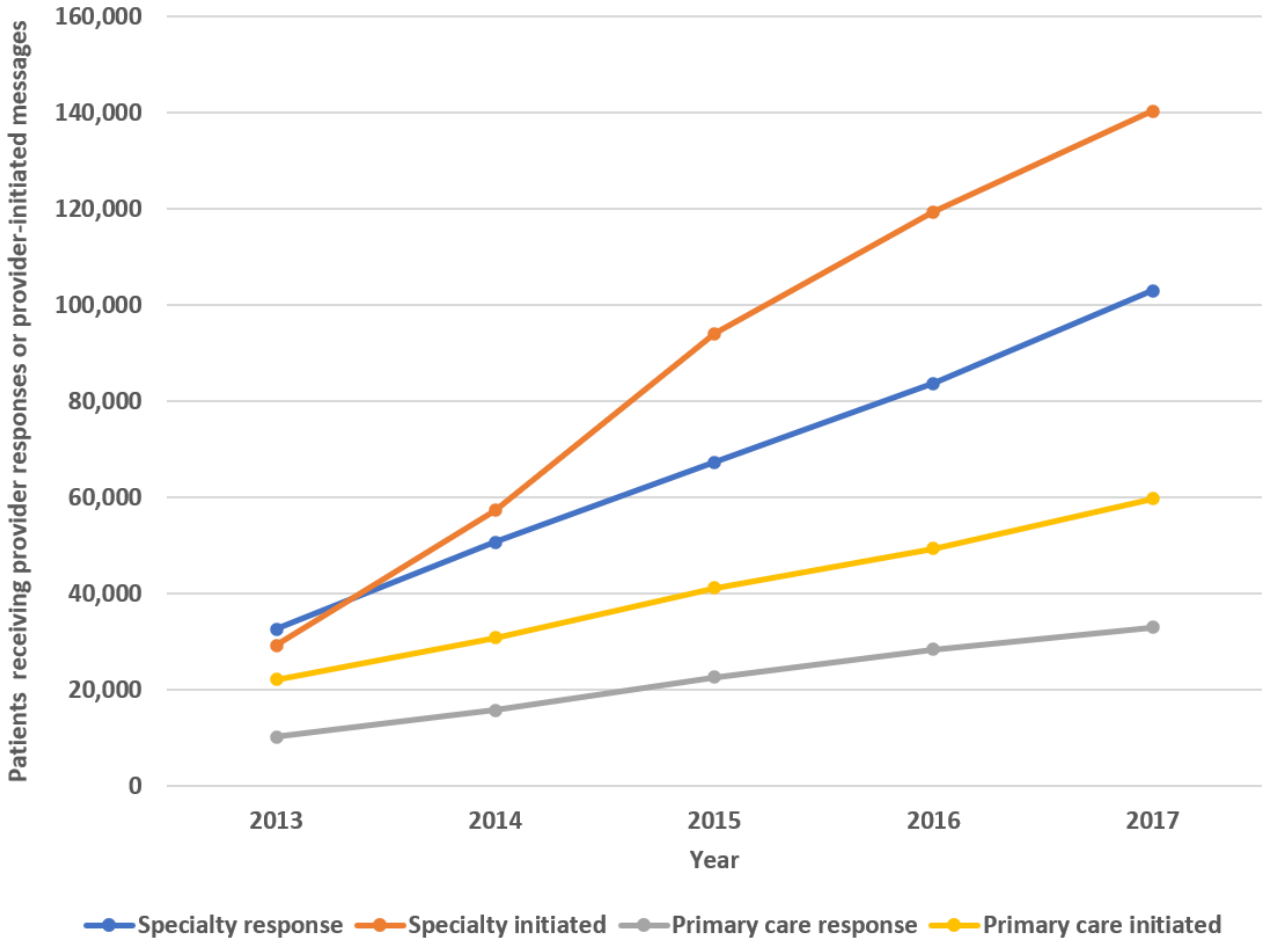


Figure 3. Message counts per unique patient by year.

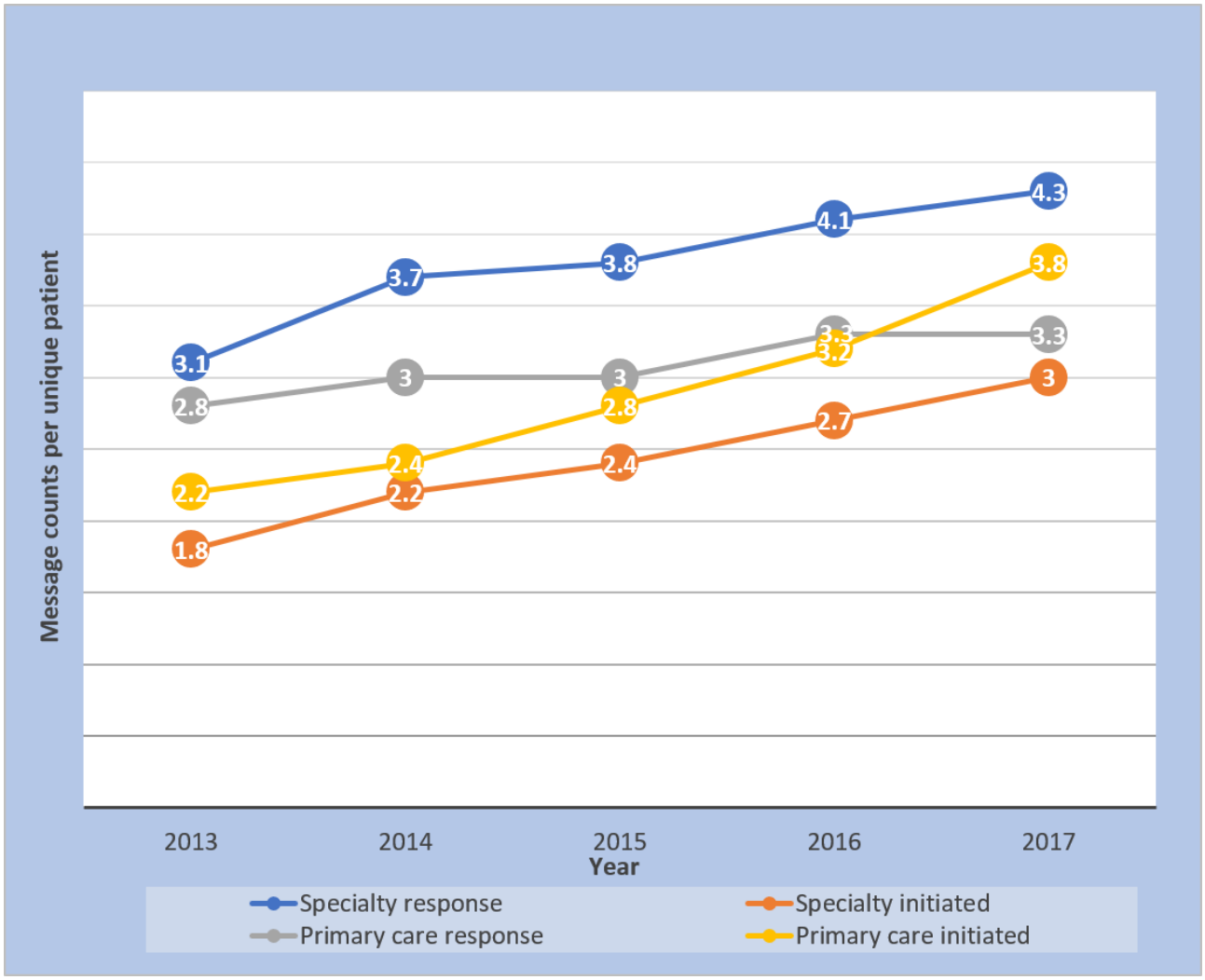


Table 2. Messages per unique patient by year.

Message type	Mean messages per unique patient by year (SD, 95% CI)					*P* value[a]
	2013	2014	2015	2016	2017	
Provider message responses from all specialty and primary care	3.6 (5.3, 3.5-3.6)	4.2 (6.6, 4.1-4.2)	4.3 (6.3, 4.3-4.3)	4.7 (6.7, 4.6-4.7)	4.9 (6.9, 4.9-4.9)	<.001
Provider-initiated messages from all specialty and primary care	2.1 (2.0, 2.1-2.1)	2.5 (2.7, 2.5-2.5)	2.8 (3.1, 2.8-2.8)	3.2 (3.6, 3.2-3.3)	3.7 (4.1, 3.7-3.7)	<.001

[a]H0: mean messages are equal across years.

Figure 4. Message responses in primary care per provider by year.

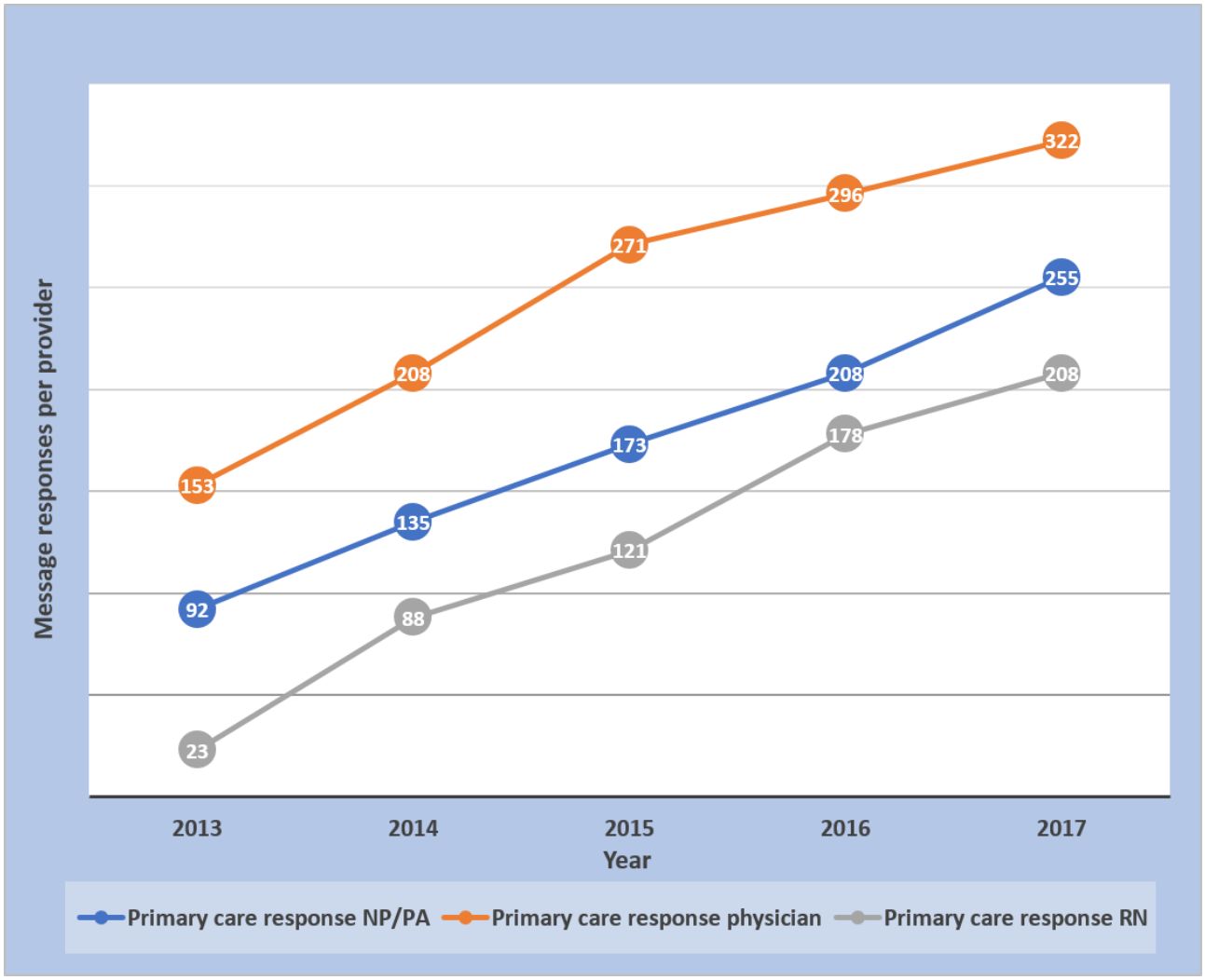
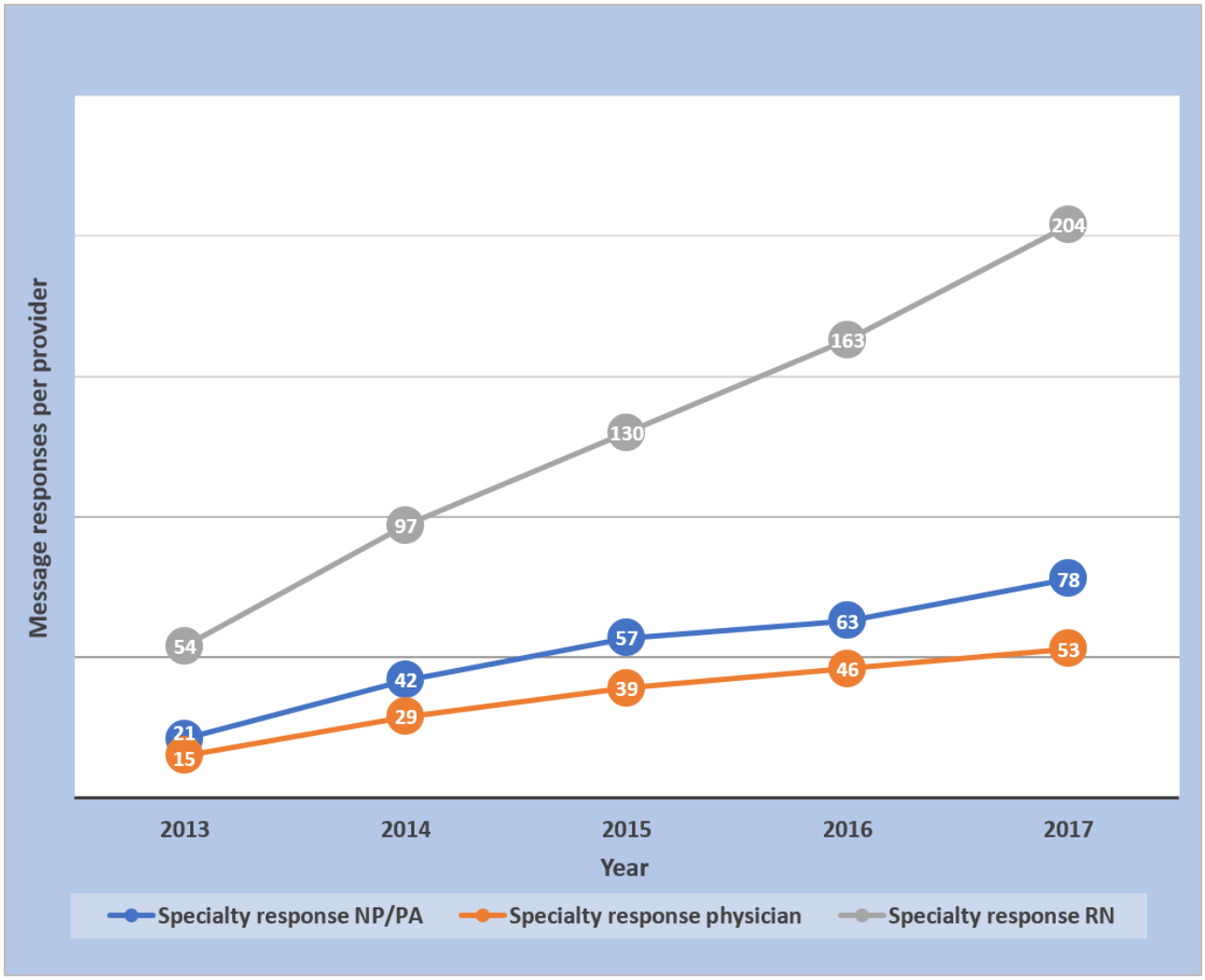


Figure 5. Message responses in specialty care per provider by year.



Tables 3 and 4 show that the number of unique patients receiving provider message responses increased from 133% to 1215% across provider groups. Some of this increase could be attributed to more patients having access to secure messages. During the same 5-year interval, there was an 88% increase in portal registration (33% registered in 2013 to 62% in 2017). In contrast, the average number of provider-response messages per unique patient increased by at most 21% across provider groups. Nurses had the largest increases in messages per provider in both primary care (804%) and specialty care (278%). The percentage of messages completed after hours increased or decreased depending on the provider group, as shown in more detail (Tables 3 and 4). However, across all provider groups, there were more messages per provider completed after hours in 2017 than in 2013 (Tables 3 and 4).

Figure 6 shows the change in work distribution for responding to messages over 5 years. Statistical analysis using the Cochran-Armitage test for trend for the data in Figure 6 showed that there were statistically significant downtrends in the percentage of message responses completed by physicians as well as *other* ($P<.001$). There were also statistically significant uptrends in the percentage of message responses completed by nurses and NP/PAs ($P<.001$).

Figure 7 shows the percentage of messages completed after hours by staff type and year. Statistical analysis using the Cochran-Armitage test for trend of the data shown in Figure 7 showed that the percentage of after-hours message responses trended up for primary care physicians and primary care NP/PAs from 2013 to 2018 (each with $P<.001$). There was no significant trend in after-hours specialty physician responses and primary care RN responses ($P=.10$ and $P=.11$, respectively). There was a significant downtrend for after-hours RN specialty responses and NP/PA specialty responses (each with $P<.001$). It should be noted that Mayo Clinic has salaried physicians, NP, and PA staff and does not base salary or any other compensation on the numbers of secure messages answered or initiated, whether during or after hours. However, for hourly compensated staff in nursing and other nonphysician/NP/PA staff, overtime work would be compensated.

Table 3. Primary care messages.

Message category and year	Message count	Unique patients	Provider count	Messages per patient	Messages per provider	Percentage completed after hours, %	Messages per provider completed after hours
Physician responses to messages							
2013	20,299	7441	133	2.73	153	18.8	29
2017	47,700	17,323	148	2.75	322	21.0	68
Change (%)	+135	+133	+11	+1	+110	+11.6	+134
NP[a]/PA[b]responses to messages							
2013	4864	2105	53	2.31	92	10.0	9
2017	25,772	11,572	101	2.23	255	19.5	50
Change (%)	+430	+450	+91	−3	+177	+94	+456
RN[c] responses to messages							
2013	2053	1261	89	1.63	23	5.0	1
2017	32,904	16,587	158	1.98	208	3.1	7
Change (%)	+1503	+1215	+78	+21	+804	−38	+600
Physician-initiated messages							
2013	36,093	17,691	133	2.04	271	N/A[d]	N/A
2017	147,020	44,049	153	3.34	961	N/A	N/A
Change (%)	+307	+149	+15	+64	+255	N/A	N/A
NP/PA-initiated messages							
2013	9095	2105	47	4.32	194	N/A	N/A
2017	54,872	9266	89	5.92	617	N/A	N/A
Change (%)	+503	+340	+89	+37	+218	N/A	N/A
RN[d]-initiated messages							
2013	2119	1063	57	1.99	37	N/A	N/A
2017	18,278	11,039	160	1.66	114	N/A	N/A
Change (%)	+763	+938	+181	−17	+208	N/A	N/A

[a]NP: nurse practitioner.

[b]PA: physician assistant.

[c]RN: registered nurse.

[d]N/A: not applicable. The percentage completed after work was not applicable in the provider-initiated messages because of the high percentage of automation.

Table 4. Specialty care messages.

Message category and year	Message count	Unique patients	Provider count	Messages per patient	Messages per provider	Percentage completed after hours, %	Messages per provider completed after hours
Physician responses to messages							
2013	19,919	7677	1337	2.59	15	21.0	3
2017	104,624	36,825	1989	2.84	53	21.4	11
Change (%)	+425	+380	+49	+10	+253	+2	+267
NP[a]/PA[b]responses to messages							
2013	4715	2103	229	2.24	21	10.9	2
2017	28,668	12,542	369	2.29	78	10.2	8
Change (%)	+508	+496	+61	+2	+271	−6	+300
RN[c]responses to messages							
2013	25,635	7966	475	3.22	54	8.5	5
2017	157,478	42,892	772	3.67	204	6.2	13
Change (%)	+514	+438	+63	+14	+278	−27	+160
Physician-initiated messages							
2013	20,711	13,982	993	1.48	21	N/A	N/A
2017	189,997	92,503	2048	2.05	93	N/A	N/A
Change (%)	+817	+562	+106	+39	+343	N/A	N/A
NP/PA-initiated messages							
2013	5867	4382	143	1.34	41	N/A	N/A
2017	49,697	28,911	344	1.72	144	N/A	N/A
Change (%)	+747	+560	+141	+28	+251	N/A	N/A
RN-initiated messages							
2013	14,616	6458	341	2.26	43	N/A	N/A
2017	99,421	38,128	732	2.61	136	N/A	N/A
Change (%)	+580	+490	+115	+15	+216	N/A	N/A

[a]NP: nurse practitioner.

[b]PA: physician assistant.

[c]RN: registered nurse.

[d]N/A: not applicable.

Figure 6. Percentage staff distribution responding to patient messages by year.

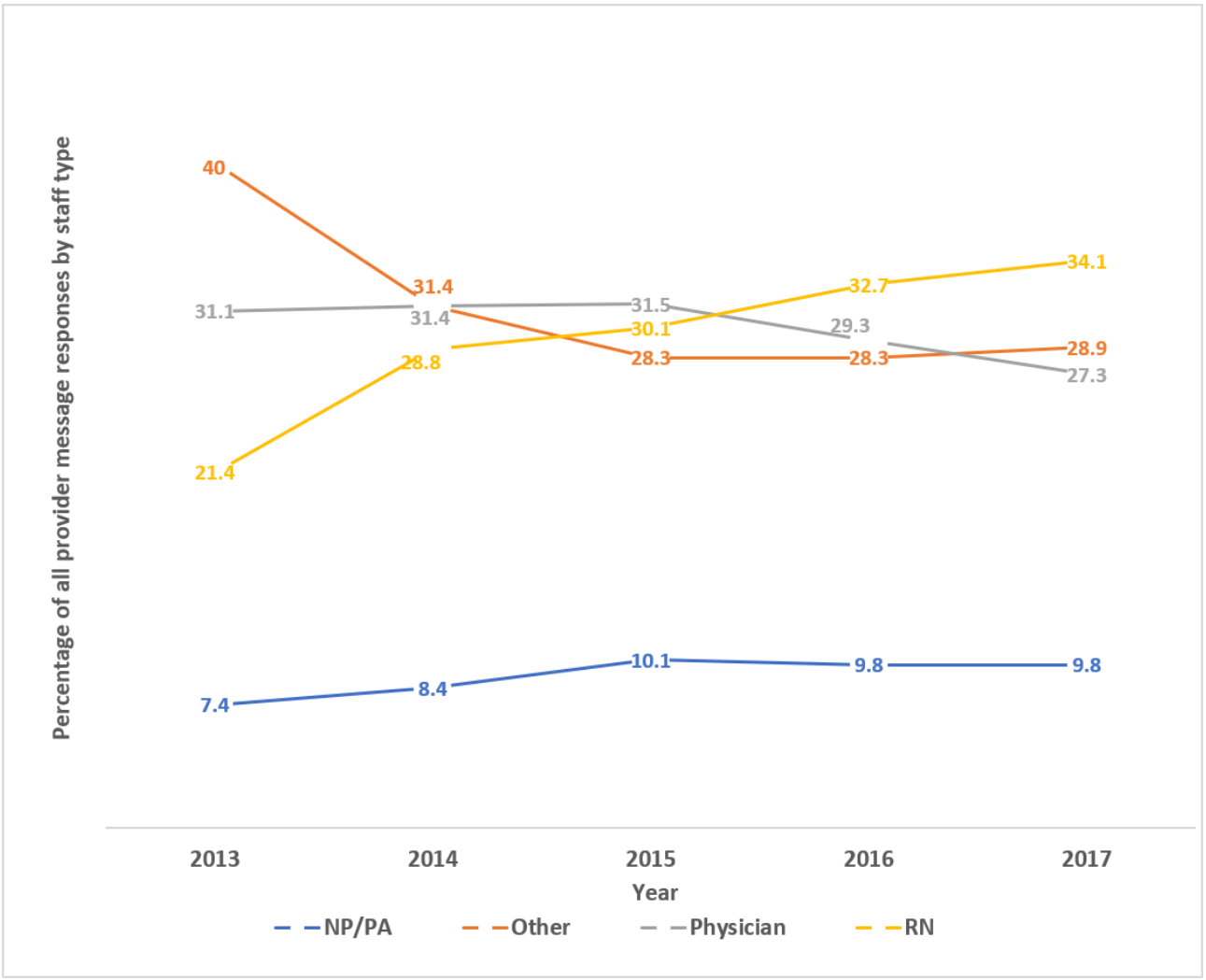
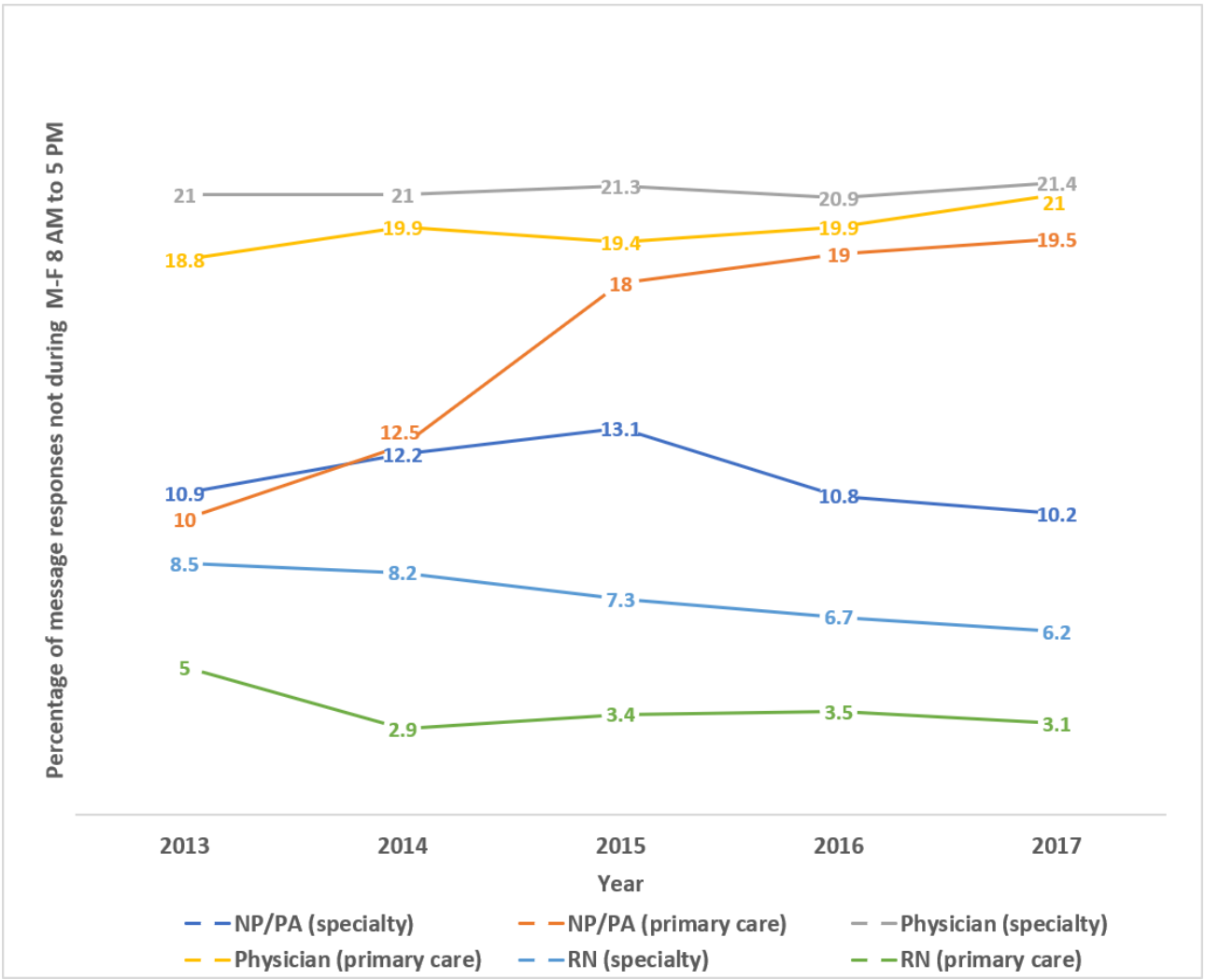


Figure 7. Percentage provider message responses outside of Monday-Friday 8 AM to 5 PM by year.



Content Abstraction and Word Counts

In the primary care provider-initiated messages, automated messages accounted for 66%, 53%, and 18% of the physician, NP/PA, and nurse-initiated messages, respectively. For the specialty practice, automated messages accounted for 54%, 47%, and 13% of the physician, NP/PA, and nurse-initiated messages, respectively.

There were only a few provider-response messages that were limited to just an acknowledgment, such as “Thanks for the update.” These acknowledgment messages accounted for only 2%, 1%, and 1% of the primary care physician, NP/PA, and nurse response messages, respectively. Similarly, for specialty providers, brief acknowledgment messages accounted for 2%, 6%, and 3% of physician, NP/PA, and nurse response messages, respectively.

Our message content review revealed that provider-response messages sometimes included a reference to an additional provider who was involved in some way with the response. For the primary care providers, 2%, 1%, and 24% of the respective physician, NP/PA, and nurse responses had evidence of involvement of another provider in the message response. With the specialty providers, other providers were involved in 1%, 6%, and 43% of the responses from physicians, NP/PA, and nurse responses, respectively. It should be noted that in our samples of 100 provider-response messages, we found no automated responses in the physician, NP/PA, or nurse messages responding to patient-initiated messages.

The median word counts from provider responses did not vary much over the course of the study; they remained just over 70 words for both the primary care and the specialty practices. Some of the word count was a standard signature that contained some packaged terms thanking patients for using the portal.

Discussion

Principal Findings

Both specialty and primary care practices experienced a large increase in the number of provider message responses as well as the number of provider-initiated messages to patients. There was no single provider category that took the brunt of the message volume increase. In fact, the provider categories of physicians, NP/PAs, RNs, and *other* all shared in handling the responses to patient messages. All these provider categories also shared in the increased volume of provider-initiated messages.

The large increase in message volume was not because of increases in patient visits. Patient visits increased by <10%,

whereas during the same time, provider responses to messages increased by 288% in the primary care practice and 345% in the specialty practice. The rise in responses to patient messages was also much greater than the 88% increase in portal registration during that time. These facts support the finding that there was increasing provider engagement in messaging during the course of the study. That is, an increasing proportion of providers were responding to and sending messages. However, the message volume was not just driven by more providers messaging their patients; there were increases in message volumes per provider across the board (Figures 4 and 5). Messages per patient also increased in both primary and specialty care practices (Figure 3 and Table 2). Hoonakker et al and Wolcott et al [7,9] reported that patient messaging was associated with providers initiating messages. Perhaps the increase in provider-initiated messages encouraged individual patients to engage more frequently in secure messages.

Implementation of secure messages was staggered, with primary care starting in 2010 and specialty practices in 2013. As a result, we had the opportunity to examine 62 months of secure message volumes for different time sequences. We observed specialty practice volumes for 62 months, starting from initiation in 2013 to 2018. As the primary care practice secure messages were implemented 3 years earlier (2010), the same 62 months encompassed years 4 to 9 of primary care message volumes. Over the same 62 months, but at different stages of experience with secure messages, secure message volumes showed continued growth both in primary care practice and in specialty practice. When separated into categories of provider message responses and provider-initiated messages, both categories showed a consistent rise over the 62 months.

Work After Work

In an ideas and opinion paper in the *Annals of Internal Medicine*, DiAngi et al [8] described some novel metrics for examining EHR use. This included a *work after work* metric that "captures the hours the clinician spends logged into the EHR during evenings, weekends, and vacations." The novel metrics also included what was termed *fair pay*, which are metrics that track "uncompensated EHR work, such as answering patient emails, providing medication refills..." [8]. Our study shows that approximately 20% of the time physicians completed message responses outside the usual business hours of 8 AM to 5 PM. There was a statistically significant uptrend over time for primary care physicians; their after-hours messaging increased from 19% to 21%, but not for specialty physicians whose after-hours messaging stayed stable between 21% and 22%. This finding is similar to that of Arndt et al [10], who found that approximately 24% of the EHR work done by physicians (1.4 hours out of 5.9) occurred after hours. NP/PAs in the specialty practice also used after-work hours for approximately 12% of their message responses across the study period. The largest work after work increase was in the NP/PA category in primary care, whose percentage of messages sent after work hours increased from 10% to 19%. The RN message responses, despite the increase in volume, had a statistically significant downtrend in percentage after-hours message completion.

As noted previously, physicians as well as the NP/PA staff are salaried, whereas a large number of nurses are on hourly wages. Although there was no statistically significant upward trend in percentage of message responses from specialty physicians completed after hours and only a couple of percentage points increase for primary care physicians, it remains that approximately 20% of these provider messages from physicians and primary care NP/PAs were sent after hours. There have been several papers associating EHR with burnout [11,12]. With the increase in secure messages, we thought that providers might be doing a larger percentage of the work after hours. Although there was a statistically significant increase in the percentage of after-hours messages completed by some provider groups, the rate of increase was low compared with the overall increase in the message volumes. However, because of the increase in message volume, all the provider groups completed more after-hours messages per provider in 2017 than in 2013 (Tables 3 and 4).

After data collection for this study was complete, Mayo Clinic switched to the Epic EHR. Epic has data collection methods that track message volumes and individual provider input (voice and keyboard) and can give management and providers feedback on the time spent in sending and receiving messages. This granular data about provider EHR activity throughout the day can be used to better identify the overall impact of messaging on provider workloads. The increase in provider secure messages shows the need for further investigations to examine the best practices in answering messages. In addition, the expanding role of secure messages needs to be considered in future studies of provider burnout.

In addition to continued examination of the work after work associated with secure messages, there is an opportunity to continue to assess how the increase in provider messaging may influence the tone of the messages. Hogan et al [13] evaluated some of the tone involved in these provider messages and noted that 25% of messages from health care team members appeared hurried. Newer informatics tools addressing sentiment analysis should help examine the content of portal messages and the sentiments associated with them [14].

Practice Implications

We found a high rate of growth of secure messages not attributable to increased patient visits, and the secure message growth continued to rise several years after implementation in primary care and specialty practices. Previous studies both at Mayo Clinic and elsewhere have demonstrated some of the effects of secure messages on subsequent face-to-face visits and some safety aspects related to secure messages [15-21], but we do not know all the ways that secure messages can affect the health care system. If secure messages cause a decrease in an equivalent volume of more time consuming, nonreimbursable telephone communication and letter correspondence, the increase in secure message volume may be a marker of increased efficiency. However, at least one study has shown no impact of messages on telephone message volume [22]. There is a need for further studies examining all forms of patient-provider communication, including the cost of *telephone tag* and transcription for letter correspondence to obtain a more

comprehensive picture of the economic impact of secure messages [23].

Our study showed that large numbers of providers and patients are engaging in secure messages. From the provider standpoint, our study shows that this includes not only physicians but also large numbers of other providers, including nurses, NP/PA staff, and other groups. Nurses had the highest increases in message responses per provider and responded to more messages per provider in the specialty practice than the specialty physician and NP/PA groups combined (Figure 5). Regarding the division of work among staff, it deserves reemphasis that our content review found 43% of the specialty nurse responses, and 24% of the primary care nurse responses had evidence of input from another provider. Laccetti et al [24], who studied cancer center secure messages, also found a sizable percentage of messages was handled by nurses (29%) and other nonphysician staff. As message volumes continue to rise, it will be important to efficiently divide the message responses among staff so that those that can be handled by nursing or other ancillary staff will not be sent to physicians. Cronin et al [25,26] at Vanderbilt have been working on automating the important job of classifying and triaging patient messages. With the participation of multiple levels of staff in messages, our study underscores the importance of further examination of how secure messages are being used and the potential importance of trainable *rules of engagement* for different staff categories responding to messages [27].

These messages represent a new avenue to access medical care. Secure messages can be more convenient than telephonic services, which often have circumscribed hours of operation and can interpose several call transfers and waits between providers, care teams, receptionists, and patients. In addition, for many institutions, including Mayo Clinic, patient secure messages are answered free of charge. Thus, for those with web-based access, secure messages can be an attractive alternative to access a provider. Hospitalized patients are also now accessing secure messages through inpatient portals. Initial studies have shown that secure messages to providers in the hospital environment have not been as highly used by patients as other inpatient portal features [28]. However, nurses and other hospital staff have seen benefits from the inpatient portal, and there is new insight on how to best introduce the inpatient portal to hospitalized patients [29,30]. Hospitalized patients who are introduced to the inpatient portal may be more likely to engage in secure messages after hospital discharge when nurses are more than a few steps away [28].

Automated messages were a large part of the provider-initiated messages; at this point, there is limited data concerning the impact of these messages and patient acceptance of them. In addition, there were provider-response messages that indicated that more than one provider was involved in the message response. It will be important to understand the work that goes into these message responses as newer forms of payment for services are being considered.

Our study shows the need to carefully examine the economics and outcomes of secure message responses in both accountable care and fee-for-service models. The increasing message volume comes in the context of a fixed number of work hours. As more time is spent with messages, there will be increasing pressure on face-to-face time, regardless of the payment system. The increase in messages is likely to cause significant outcomes and economic impacts with either payment model.

Limitations

This study had several limitations. First, this was limited to one large multispecialty group located in North Central United States. The patients were mostly white and well educated, and the percentage of patients seen in the clinic who are registered on the portal increased to 62% over the course of the study. Our study also had a much higher rate of provider-initiated messages compared with a veteran's health administration study by Shimada et al [31], which showed that only 5.5% of messages were initiated by providers proactively reaching out to patients.

As seen in Table 1, there were changes in demographics of patients who were sending messages from the initiation of the study (early users of secure messages) to 5 years later. We did not perform an in-depth analysis of the demographics of face-to-face Mayo Clinic patients simultaneously at the same time points. It is possible that some of the longitudinal shift in demographics that we saw in patients using messages could be confounded by a 5-year demographic shift of all Mayo Clinic patients.

Our experience at Mayo Clinic in Rochester, Minnesota, may not be generalizable to other multispecialty groups. Mayo Clinic has a specialty care focus, and patients both nationally and internationally come to Mayo Clinic to receive highly specialized care. As seen in Table 1, approximately 40% of specialty response messages went to patients who did not live in Minnesota. The geographical distance of many of the Mayo specialty patients may encourage a higher use of secure messages compared with other multispecialty groups whose patients have fewer geographic barriers to face-to-face visits.

LPNs were also involved in some message responses and initiated messages but accounted for only 0.7% (12,680) of message responses and 1.6% (29,774) of initiated messages, so they were put in the *other* category. The staff responding to and initiating messages in the *other* category were generally secretaries and clinical assistants involved in scheduling; some of these messages did not have an associated identifier that we could assign to either primary care or specialty care. In addition, there were small percentages of nurses and NP/PAs who transferred from specialty to primary care or vice versa; our administrative data could not correct for these individual changes. However, these changes were likely limited to the RN and NP/PA groups, as physicians were constrained by physician specialty board certification.

Conclusions

In the first 8 years of secure message use in a large multispecialty group, secure message volumes showed large increases both in response to patient messages and provider-initiated secure messages. Both primary care practices and specialty practices saw large growth rates in total messages and messages per provider. Physicians, NP/PAs, and RNs all shared in responding to the increased volume of messages.

Messages per unique patient also showed a significant increase over 5 years. The percentage of message responses after hours stayed close to 20% each year over 5 years for physicians in primary care and specialty practices.

Authors' Contributions

FN contributed to the study conception. FN, KL, EM, and TM contributed to the study design. FN, TM, and EM analyzed and interpreted the data and contributed to statistics. FN drafted the manuscript. FN, KL, TM, EM, ST, EN, and JP edited, critically revised, and approved the final version of the paper.

Conflicts of Interest

None declared.

References

1. National Cancer Institute. Health Information National Trends Survey: HINTS. 2013. Health Information Trends Survey Responses to Question 'In the Last 12 Months, Have You Used the Internet for Any of the Following Reasons' Used E-mail or the Internet to Communicate With a Doctor or Doctor's Office URL: https://hints.cancer.gov/view-questions-topics/question-details.aspx?PK_Cycle=6&qid=761 [accessed 2020-05-05]
2. Lee JL, Choudhry NK, Wu AW, Matlin OS, Brennan TA, Shrank WH. Patient use of email, Facebook, and physician websites to communicate with physicians: a national online survey of retail pharmacy users. J Gen Intern Med 2016 Jan;31(1):45-51 [FREE Full text] [doi: 10.1007/s11606-015-3374-7] [Medline: 26105675]
3. Crotty BH, Tamrat Y, Mostaghimi A, Safran C, Landon BE. Patient-to-physician messaging: volume nearly tripled as more patients joined system, but per capita rate plateaued. Health Aff (Millwood) 2014 Oct;33(10):1817-1822 [FREE Full text] [doi: 10.1377/hlthaff.2013.1145] [Medline: 25288428]
4. Cronin RM, Davis SE, Shenson JA, Chen Q, Rosenbloom ST, Jackson GP. Growth of secure messaging through a patient portal as a form of outpatient interaction across clinical specialties. Appl Clin Inform 2015;6(2):288-304 [FREE Full text] [doi: 10.4338/ACI-2014-12-RA-0117] [Medline: 26171076]
5. Masterman M, Cronin R, Davis S, Shenson J, Jackson G. Adoption of secure messaging in a patient portal across pediatric specialties. AMIA Annu Symp Proc 2016;2016:1930-1939 [FREE Full text] [Medline: 28269952]
6. Shenson JA, Cronin RM, Davis SE, Chen Q, Jackson GP. Rapid growth in surgeons' use of secure messaging in a patient portal. Surg Endosc 2016 Apr;30(4):1432-1440 [FREE Full text] [doi: 10.1007/s00464-015-4347-y] [Medline: 26123340]
7. Hoonakker P, Carayon P, Cartmill R. The impact of secure messaging on workflow in primary care: results of a multiple-case, multiple-method study. Int J Med Inform 2017 Apr;100:63-76. [doi: 10.1016/j.ijmedinf.2017.01.004] [Medline: 28241939]
8. DiAngi YT, Lee TC, Sinsky CA, Bohman BD, Sharp CD. Novel metrics for improving professional fulfillment. Ann Intern Med 2017 Nov 21;167(10):740-741. [doi: 10.7326/M17-0658] [Medline: 29052698]
9. Wolcott V, Agarwal R, Nelson DA. Is provider secure messaging associated with patient messaging behavior? Evidence from the US army. J Med Internet Res 2017 Apr 6;19(4):e103 [FREE Full text] [doi: 10.2196/jmir.6804] [Medline: 28385681]
10. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
11. Babbott S, Manwell LB, Brown R, Montague E, Williams E, Schwartz M, et al. Electronic medical records and physician stress in primary care: results from the MEMO study. J Am Med Inform Assoc 2014 Feb;21(e1):e100-e106 [FREE Full text] [doi: 10.1136/amiajnl-2013-001875] [Medline: 24005796]
12. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
13. Hogan TP, Luger TM, Volkman JE, Rocheleau M, Mueller N, Barker AM, et al. Patient centeredness in electronic communication: evaluation of patient-to-health care team secure messaging. J Med Internet Res 2018 Mar 8;20(3):e82 [FREE Full text] [doi: 10.2196/jmir.8801] [Medline: 29519774]
14. Han H, Zhang Y, Zhang J, Yang J, Zou X. Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias. PLoS One 2018;13(8):e0202523 [FREE Full text] [doi: 10.1371/journal.pone.0202523] [Medline: 30142154]
15. Markle Fondation. 2003. Americans Want Benefits of Personal Health Records URL: http://www.markle.org/publications/950-americans-want-benefits-personal-health-records [accessed 2020-04-15]
16. North F, Crane SJ, Chaudhry R, Ebbert JO, Ytterberg K, Tulledge-Scheitel SM, et al. Impact of patient portal secure messages and electronic visits on adult primary care office visits. Telemed J E Health 2014 Mar;20(3):192-198 [FREE Full text] [doi: 10.1089/tmj.2013.0097] [Medline: 24350803]

17. North F, Crane SJ, Stroebel RJ, Cha SS, Edell ES, Tulledge-Scheitel SM. Patient-generated secure messages and eVisits on a patient portal: are patients at risk? J Am Med Inform Assoc 2013;20(6):1143-1149 [FREE Full text] [doi: 10.1136/amiajnl-2012-001208] [Medline: 23703826]
18. Palen TE, Ross C, Powers JD, Xu S. Association of online patient access to clinicians and medical records with use of clinical services. J Am Med Assoc 2012 Nov 21;308(19):2012-2019. [doi: 10.1001/jama.2012.14126] [Medline: 23168824]
19. Reed M, Graetz I, Gordon N, Fung V. Patient-initiated e-mails to providers: associations with out-of-pocket visit costs, and impact on care-seeking and health. Am J Manag Care 2015 Dec 1;21(12):e632-e639 [FREE Full text] [Medline: 26760425]
20. Zhou Y, Garrido T, Chin H, Wiesenthal A, Liang L. Patient access to an electronic health record with secure messaging: impact on primary care utilization. Am J Manag Care 2007 Jul;13(7):418-424 [FREE Full text] [Medline: 17620037]
21. Baer D. Patient-physician e-mail communication: the kaiser permanente experience. J Oncol Pract 2011 Jul;7(4):230-233 [FREE Full text] [doi: 10.1200/JOP.2011.000323] [Medline: 22043186]
22. Bergmo TS, Kummervold PE, Gammon D, Dahl LB. Electronic patient-provider communication: will it offset office visits and telephone consultations in primary care? Int J Med Inform 2005 Sep;74(9):705-710. [doi: 10.1016/j.ijmedinf.2005.06.002] [Medline: 16095961]
23. Goldzweig CL, Orshansky G, Paige NM, Towfigh AA, Haggstrom DA, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013 Nov 19;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
24. Laccetti AL, Chen B, Cai J, Gates S, Xie Y, Lee SJ, et al. Increase in cancer center staff effort related to electronic patient portal use. J Oncol Pract 2016 Dec;12(12):e981-e990 [FREE Full text] [doi: 10.1200/JOP.2016.011817] [Medline: 27601511]
25. Cronin R, Fabbri D, Denny J, Jackson G. Automated classification of consumer health information needs in patient portal messages. AMIA Annu Symp Proc 2015;2015:1861-1870 [FREE Full text] [Medline: 26958285]
26. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. Int J Med Inform 2017 Sep;105:110-120 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.06.004] [Medline: 28750904]
27. Sieck CJ, Hefner JL, Schnierle J, Florian H, Agarwal A, Rundell K, et al. The rules of engagement: perspectives on secure messaging from experienced ambulatory patient portal users. JMIR Med Inform 2017 Jul 4;5(3):e13 [FREE Full text] [doi: 10.2196/medinform.7516] [Medline: 28676467]
28. Huerta T, Fareed N, Hefner JL, Sieck CJ, Swoboda C, Taylor R, et al. Patient engagement as measured by inpatient portal use: methodology for log file analysis. J Med Internet Res 2019 Mar 25;21(3):e10957 [FREE Full text] [doi: 10.2196/10957] [Medline: 30907733]
29. Hefner J, Sieck C, McAlearney A. Training to optimize collaborative use of an inpatient portal. Appl Clin Inform 2018 Jul;9(3):558-564 [FREE Full text] [doi: 10.1055/s-0038-1666993] [Medline: 30045386]
30. Hefner JL, Sieck CJ, Walker DM, Huerta TR, McAlearney AS. System-wide inpatient portal implementation: survey of health care team perceptions. JMIR Med Inform 2017 Sep 14;5(3):e31 [FREE Full text] [doi: 10.2196/medinform.7707] [Medline: 28912115]
31. Shimada SL, Petrakis BA, Rothendler JA, Zirkle M, Zhao S, Feng H, et al. An analysis of patient-provider secure messaging at two veterans health administration medical centers: message content and resolution through secure messaging. J Am Med Inform Assoc 2017 Sep 1;24(5):942-949. [doi: 10.1093/jamia/ocx021] [Medline: 28371896]

Abbreviations

EHR: electronic health record
LPN: licensed practical nurse
NP: nurse practitioner
PA: physician assistant
POS: Patient Online Services
RN: registered nurse

Edited by J Hefner; submitted 24.11.19; peer-reviewed by N Drimer, A Klingberg, T Palen; comments to author 16.02.20; revised version received 17.03.20; accepted 15.04.20; published 08.07.20.

Please cite as:
North F, Luhman KE, Mallmann EA, Mallmann TJ, Tulledge-Scheitel SM, North EJ, Pecina JL
A Retrospective Analysis of Provider-to-Patient Secure Messages: How Much Are They Increasing, Who Is Doing the Work, and Is the Work Happening After Hours?
JMIR Med Inform 2020;8(7):e16521
URL: https://medinform.jmir.org/2020/7/e16521
doi:10.2196/16521
PMID:32673238

Original Paper

The Kaiser Permanente Northern California Adult Alcohol Registry, an Electronic Health Records-Based Registry of Patients With Alcohol Problems: Development and Implementation

Vanessa A Palzes[1], MPH; Constance Weisner[1,2], MSW, DrPH; Felicia W Chi[1], MPH; Andrea H Kline-Simon[1], MSc; Derek D Satre[1,2], PhD; Matthew E Hirschtritt[1,2,3], MPH, MD; Murtuza Ghadiali[4,5], MD; Stacy Sterling[1], MSW, DrPH

[1]Division of Research, Kaiser Permanente Northern California, Oakland, CA, United States

[2]Department of Psychiatry, Weill Institute of Neurosciences, University of California, San Francisco, CA, United States

[3]Department of Psychiatry, Kaiser Permanente East Bay, Oakland, CA, United States

[4]Department of Addiction Medicine, Kaiser Permanente San Francisco Medical Center, San Francisco, CA, United States

[5]Department of Addiction Psychiatry, University of California, San Francisco, CA, United States

Corresponding Author:
Vanessa A Palzes, MPH
Division of Research
Kaiser Permanente Northern California
2000 Broadway
Oakland, CA, 94612
United States
Phone: 1 510 891 3743
Email: vanessa.a.palzes@kp.org

Abstract

Background: Electronic health record (EHR)–based disease registries have aided health care professionals and researchers in increasing their understanding of chronic illnesses, including identifying patients with (or at risk of developing) conditions and tracking treatment progress and recovery. Despite excessive alcohol use being a major contributor to the global burden of disease and disability, no registries of alcohol problems exist. EHR-based data in Kaiser Permanente Northern California (KPNC), an integrated health system that conducts systematic alcohol screening, which provides specialty addiction medicine treatment internally and has a membership of over 4 million members that are highly representative of the US population with access to care, provide a unique opportunity to develop such a registry.

Objective: Our objectives were to describe the development and implementation of a protocol for assembling the KPNC Adult Alcohol Registry, which may be useful to other researchers and health systems, and to characterize the registry cohort descriptively, including underlying health conditions.

Methods: Inclusion criteria were adult members with unhealthy alcohol use (using National Institute on Alcohol Abuse and Alcoholism guidelines), an alcohol use disorder (AUD) diagnosis, or an alcohol-related health problem between June 1, 2013, and May 31, 2019. We extracted patients' longitudinal, multidimensional EHR data from 1 year before their date of eligibility through May 31, 2019, and conducted descriptive analyses.

Results: We identified 723,604 adult patients who met the registry inclusion criteria at any time during the study period: 631,780 with unhealthy alcohol use, 143,690 with an AUD diagnosis, and 18,985 with an alcohol-related health problem. We identified 65,064 patients who met two or more criteria. Of the 4,973,195 adult patients with at least one encounter with the health system during the study period, the prevalence of unhealthy alcohol use was 13% (631,780/4,973,195), the prevalence of AUD diagnoses was 3% (143,690/4,973,195), and the prevalence of alcohol-related health problems was 0.4% (18,985/4,973,195). The registry cohort was 60% male (n=432,847) and 41% non-White (n=295,998) and had a median age of 41 years (IQR=27). About 48% (n=346,408) had a chronic medical condition, 18% (n=130,031) had a mental health condition, and 4% (n=30,429) had a drug use disorder diagnosis.

Conclusions: We demonstrated that EHR-based data collected during clinical care within an integrated health system could be leveraged to develop a registry of patients with alcohol problems that is flexible and can be easily updated. The registry's comprehensive patient-level data over multiyear periods provides a strong foundation for robust research addressing critical

public health questions related to the full course and spectrum of alcohol problems, including recovery, which would complement other methods used in alcohol research (eg, population-based surveys, clinical trials).

(JMIR Med Inform 2020;8(7):e19081) doi:10.2196/19081

KEYWORDS

electronic health records; alcohol; registry; unhealthy alcohol use; alcohol use disorder; recovery; secondary data

Introduction

Electronic health records (EHRs) provide a platform to study many diseases and health-related issues longitudinally in diverse populations, including identification of patients with (or at-risk of developing) conditions, and tracking treatment progress and recovery. The development of EHR-based disease registries has aided health care professionals and researchers in increasing their understanding of chronic illnesses and how to manage them [1]. For example, disease registries can facilitate the coordination of care within a health system [2,3]. However, they can also enable research on treatment effectiveness and patient outcomes, complementing other methods that are costly for repeated data collection in large populations (eg, clinical trials, surveys) [4].

Despite excessive alcohol use being a significant contributor to the global burden of disease and disability [5], to our knowledge, no population-, health system–, or EHR-based registries of individuals with alcohol problems exist. In 2016, excessive alcohol use accounted for 3 million deaths worldwide (5.3% of all deaths), which was higher than that of common conditions, such as diabetes (2.8%), road injuries (2.5%), tuberculosis (2.3%), and hypertension (1.6%) [5]. Alcohol-related death rates in the United States have also increased substantially over the past decade [6], accelerating over recent years [7]. Alcohol use is a known risk factor for serious medical conditions, including pancreatitis [8], stroke [9], and breast cancer [10], and can lead to alcohol use disorder (AUD) and alcoholic liver cirrhosis [11]. Alcohol can also impact the course of disease progression, management, and treatment outcomes for a range of conditions, including diabetes [12], depression [13,14], and anxiety [15]. While the prevalence of excessive alcohol use in the general US population ranges from 6% to 28% (depending on the definition and whether individuals with an AUD diagnosis are included) [16,17], there is evidence that it is increasing [17,18]. Therefore, alcohol problems are a significant public health concern that would benefit from being the primary focus of a disease registry.

The goal of the overall project was to assemble a registry of patients with alcohol problems by leveraging comprehensive EHR-based data within Kaiser Permanente Northern California (KPNC). The registry was developed for research specifically, but with the potential for future clinical or administrative applications such as quality improvement. KPNC is an integrated health care system that provides primary and specialty care internally (including addiction medicine and psychiatry). It has a mature, fully developed Epic EHR system (Epic Systems, Verona, WI), Kaiser Permanente (KP) HealthConnect, that stores data collected throughout the full course of patient care since 2005. Additionally, KPNC has conducted over 12 million alcohol screenings among 4 million adult members since June 2013 as part of a systematic alcohol screening, brief intervention, and referral to treatment initiative in primary care [19], which adds a robust patient-reported element to clinical data recorded in the EHR. Therefore, longitudinal, multidimensional patient-level data can be obtained (including alcohol use, health service utilization, diagnoses, medications, laboratory tests, and responses to health questionnaires), providing a unique opportunity to study the onset and progression of alcohol problems, care provided during all phases (ie, follow-up, management, continuity of care), and measurable outcomes such as changes in drinking. The objective of this paper was to describe the protocol used to develop the registry and to characterize patients who met eligibility criteria, including underlying health conditions. We include our methodological approach and considerations related to the registry, which we hope will be useful to other research teams and health systems with the ability to track unhealthy alcohol use, AUDs, and alcohol-related health problems (ie, conditions that are entirely attributable to alcohol).

Methods

Setting

KPNC serves 4.3 million members, comprising about one-third of the population in Northern California. The membership is diverse and highly representative of the US population with access to care [20]. Membership includes enrollees from Medicaid (12%), Medicare (16%), employer-based plans, and health insurance exchanges. KPNC members have direct access to specialty care clinics, including addiction medicine and psychiatry [21].

In June 2013, KPNC implemented Alcohol as a Vital Sign, a systematic alcohol screening, brief intervention, and referral to treatment initiative, in adult primary care [19]. While the initiative is primary care-based, the EHR screening tools are available for use in outpatient medical departments. KPNC has maintained an average 87% screening rate systemwide in adult primary care. As part of the screening, patients are asked three questions about their alcohol use, including a modified version of the evidence-based National Institute on Alcohol Abuse and Alcoholism (NIAAA) single-item screening question [16] (tailored to the patient's age and sex)—"How many times in the past three months have you had 5 or more drinks containing alcohol in a day?" (for men aged 18-65 years), or "4 or more drinks" (for all women and for men aged ≥66 years)—and two questions that are used to calculate average drinks consumed per week—"On average, how many days per week do you have an alcoholic drink?" and "On a typical drinking day, how many drinks do you have?" The EHR issues a best practice alert during a primary care visit when screening is required (ie, first visit,

annually, or every six months if unhealthy alcohol use was previously reported). The medical assistant may skip these questions for a variety of reasons (eg, late appointment arrivals, forgetting), and patients may decline to respond.

Data Sources

Registry data are leveraged from two existing data sources: Clarity (GridApp Systems, Inc) and the Virtual Data Warehouse (VDW). Clarity is the back-end database of EHR data collected in KP HealthConnect, which we used to extract alcohol screening data. The VDW is a distributed data model developed by the Health Care Systems Research Network (HCSRN) to maintain single extract, transform, and load processes that efficiently create relational tables useful for research [22]. The VDW gathers data from various EHR-based sources, including Clarity, and legacy systems that were used before the implementation of KP HealthConnect in 2005. The VDW data has been developed over many years with standardized data definitions and formats, and rigorous quality assurance.

Objective and Aims

In collaboration with NIAAA, we defined the objective of the registry and target population and formed research aims to frame the registry's scope. The purpose of the registry is to study the full course of alcohol problems with the flexibility to address many research questions, such as the escalation of unhealthy drinking to development of AUDs and alcohol-related health problems, and the ability to be updated with new data. The target population for the registry is adult patients diagnosed with an alcohol problem and those at risk of developing one.

Protocol

We developed a protocol for building the registry (available upon request), following recommendations from the US Agency for Healthcare Research and Quality [4] and other disease registries [23,24], and received approval by the Institutional Review Board at KPNC. We benchmarked our approach to that of other disease registries at KPNC (eg, HIV [25], diabetes [26], cancer [27], opioid use [28]), to determine feasibility, data storage, and access. We surveyed the literature and involved KPNC physicians in psychiatry and addiction medicine to help select key data elements and clarify data definitions. We established a plan for leveraging available data by characterizing eligibility criteria for inclusion, defining the structure of the registry, and identifying core data elements and variables needed to address the research aims. We developed codebooks to define the scope of the registry (eg, diagnosis codebook of International Classification of Diseases, 9th Revision, Clinical Modification [ICD-9] and 10th Revision, Clinical Modification [ICD-10] codes), which can easily be updated to extend the breadth of data that the registry captures.

Inclusion Criteria

We included adult patients (age ≥18 years) with unhealthy alcohol use, an active AUD diagnosis, or an alcohol-related health problem, from any department or encounter setting within the health system. The initial registry cohort includes patients who met these criteria between June 1, 2013, (when Alcohol as a Vital Sign was implemented) to May 31, 2019. The patient's index date was the first date in which the patient met eligibility criteria during the study period.

Unhealthy alcohol use was identified using systematic alcohol screening data collected as part of Alcohol as a Vital Sign. Using NIAAA recommended drinking guidelines [16], we defined unhealthy alcohol use as exceeding either the daily (≥5 drinks/day for men aged 18-65 years, or ≥4 drinks/day for women and for men aged ≥66 years) or weekly (>14 drinks/week for men aged 18-65 years, or >7 drinks/week for women and for men aged ≥66 years) drinking limit. To determine which risk threshold to use, we used the patient's age and EHR-assigned sex, which is directly provided by the purchaser of a health insurance plan during enrollment. For patients with unknown sex (n=270), we used their sex assigned at birth (n=45), if available, which is a patient-reported variable collected along with gender identity in the EHR. Otherwise, we imputed sex based on the patient's age and which single-item screening question was asked (n=225). If the patient was aged 18-65 years and asked, "How many times in the past three months have you had 5 or more drinks containing alcohol in a day?" sex was imputed as male (n=106), otherwise as female (n=119).

ICD-9 and ICD-10 codes given at any encounter at KPNC or through a claim were used to identify patients with a diagnosis of an active AUD (excluding remission codes) or an alcohol-related health problem (Table 1) [29].

Table 1. International Classification of Diseases (ICD) codes for identification of active alcohol use disorders and alcohol-related health problems as part of inclusion criteria for the Kaiser Permanente Northern California Adult Alcohol Registry.

Disorder, ICD[a] version, and code	Description
Alcohol use disorders	
ICD-9	
291[b]	Alcohol-induced mental disorders (eg, alcohol withdrawal delirium)
303[b], except 303.03 and 303.93[c]	Alcohol dependence syndrome
305.0[b], except 305.03[c]	Nondependent alcohol abuse
ICD-10	
F10.9[b]	Alcohol use, unspecified (includes alcohol-induced mental disorders)
F10.2[b], except F10.21[d]	Alcohol dependence
F10.1[b], except F10.11[d]	Alcohol abuse
Alcohol-related health problems	
ICD-9	
357.5	Alcoholic polyneuropathy
425.5	Alcoholic cardiomyopathy
535.3[b]	Alcoholic gastritis
571.0-571.3	Alcoholic liver disease
ICD-10	
G31.2	Degeneration of nervous system due to alcohol
G62.1	Alcoholic polyneuropathy
G72.1	Alcoholic myopathy
I42.6	Alcoholic cardiomyopathy
K29.2[b]	Alcoholic gastritis
K70[b]	Alcoholic liver disease
K86.0	Alcohol-induced chronic pancreatitis

[a]ICD: International Classification of Diseases.

[b]Any (or no) additional digits.

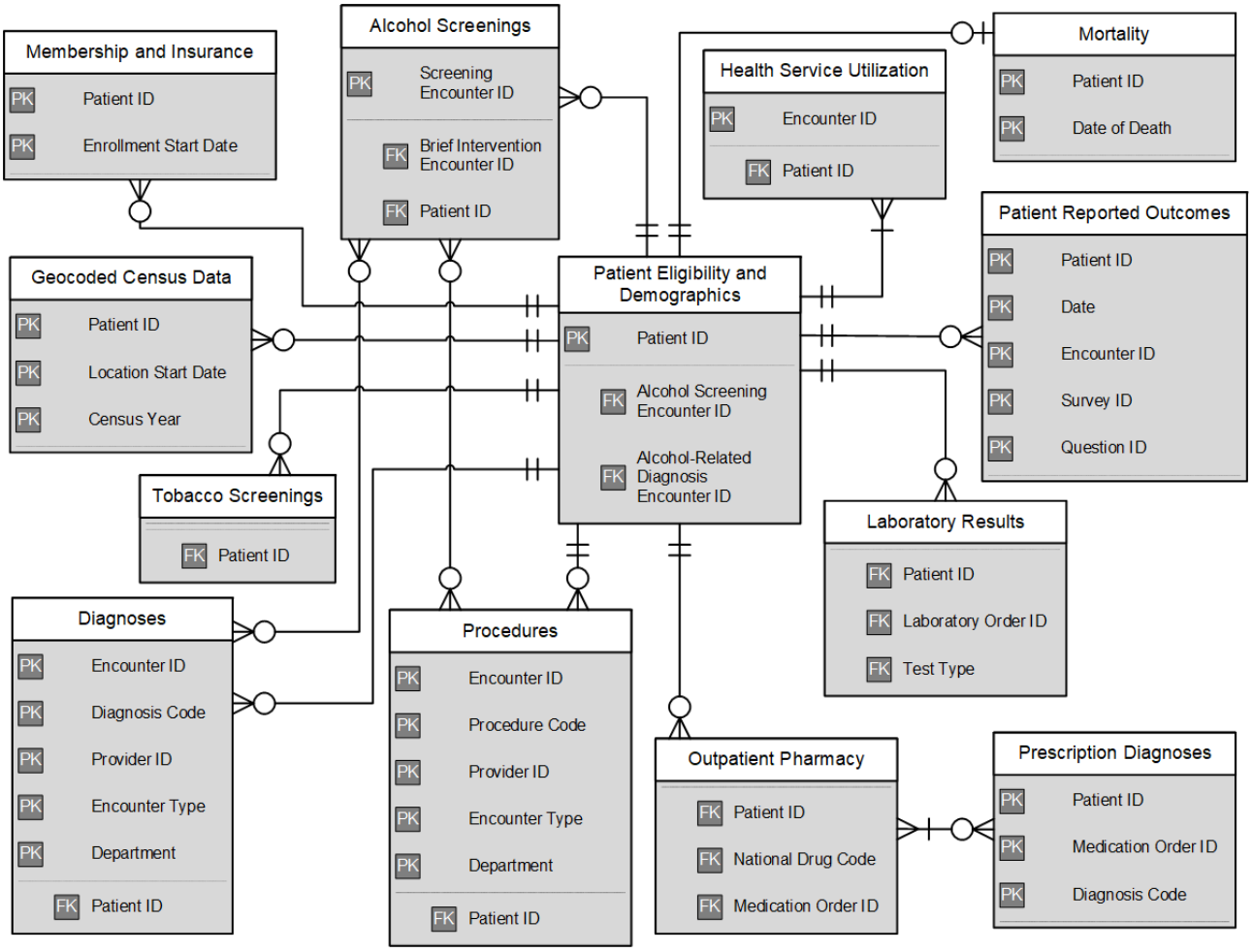
[c]303.03, 303.93, and 305.03 are ICD-9 remission codes.

[d]F10.21 and F10.11 are ICD-10 remission codes.

Structure and Data Elements

Like the VDW [22], the registry was designed as a distributed data model where each file contains one main content area, and files can be linked through key variables (eg, person ID, encounter ID; Figure 1). Main content areas include patient eligibility and demographics, alcohol screenings, membership and insurance, geocoded census data, diagnoses, procedures, outpatient pharmacy, prescription diagnoses, laboratory results, patient-reported outcomes, tobacco screenings, health service utilization, mortality, and total KPNC membership. More detailed descriptions of the data elements can be found in Multimedia Appendix 1, and specific diagnoses tracked in the registry in Multimedia Appendix 2. In each file, we retained and created variables necessary to address our research aims and used codebooks to filter the data efficiently (available upon request). We included all data from 1 year prior to the patient's index date (serving as a time window for identifying co-occurring health conditions [30]) through the end of the study period.

Figure 1. Entity-relationship diagram representing the data structure of core files in the Kaiser Permanente Northern California Adult Alcohol Registry. Primary key variables are unique identifiers that can be used with foreign key variables to link data across files. PK: primary key; FK: foreign key.



Implementation

Implementation of the protocol took about 10 months with 50% programmer time effort. We wrote programs using SAS software, version 9.4 of the SAS System for Unix (SAS Institute), to build the registry, which were designed to minimize user interaction and could be used again to refresh the registry data (eg, using macros and macro variables). We minimized data cleaning to allow future studies to make their own decisions regarding the use of the data. We created a data dictionary to describe the files and variables that comprise the registry. We also developed queries for quality control, such as identifying missing data and characterizing data storage requirements. Last, we created reporting tools to display trends of the registry data over time.

Maintenance

Since the EHR is a constantly changing data environment, refreshing the registry with new data requires programs and documentation to be updated. For example, source variables and tables may be renamed or become deprecated during upgrades of data systems. The amount of time required to refresh the registry depends on the quantity and types of changes needed (eg, adding more ICD codes versus editing SAS programs), but may take anywhere from an hour to several days. Receiving ongoing feedback of the registry as research staff use it for their projects is also critical to ensuring the registry's validity and usefulness.

Analysis of the Registry Cohort

We calculated the prevalence of alcohol problems among all adult KPNC patients who had at least one encounter with the health system between June 1, 2013, and May 31, 2019. We conducted descriptive analyses to describe demographic, clinical (eg, medical and mental health conditions), and insurance characteristics of the registry cohort. We included only key variables in the current analysis to compare the registry cohort to those in other published studies. All characteristics, such as age, were based on the patient's index date. We estimated patients' household income and education using US Census data that has been geocoded to patients' closest residential addresses in the year prior to and including the month of their index date. If the index date was before January 1, 2017, we used the 2010 US Census data; otherwise, we used 2017 data, since census block boundaries can change over time [31]. We used the median household income of the census block to estimate patients' household income and categorized patients into groups used in prior epidemiologic studies of the general US population [18,32]. The education level with the highest proportion of households in the census block was used to estimate patients' education. To identify smoking status, we used the closest screening in the year prior to and including the

month of the index date. We calculated the Charlson comorbidity score, which estimates the 1-year mortality risk based on a weighted score of 17 medical conditions [33], and identified chronic medical and mental health conditions and substance use disorder diagnoses in the year prior to the index date. All analyses were conducted using SAS software version 9.4.

Results

We identified 723,604 adult patients eligible for inclusion in the registry between June 1, 2013, to May 31, 2019: 631,780 with unhealthy alcohol use, 143,690 with an AUD diagnosis, and 18,985 with an alcohol-related health problem, anytime during the study period. Counts are not independent, as 65,064 patients met two or more eligibility criteria. Of 4,973,195 adult KPNC patients with at least one encounter with the health system during the study period, the prevalence of unhealthy alcohol use was 13% (631,780/4,973,195), the prevalence of AUD diagnoses was 3% (143,690/4,973,195), and the prevalence of alcohol-related health problems was 0.4% (18,985/4,973,195).

The registry cohort was about 60% (n=432,847) male and 40% (n=290,755) female, and there were 2 patients with other/unknown sex. In regard to gender, 0.1% (n=688) of the cohort were gender minorities (transgender, nonbinary, or other gender). The median age was 41 years (IQR=27; Table 2). The cohort was 19% (n=138,925) Latino/Hispanic, 11% (n=76,197) Asian, Native Hawaiian or Pacific Islander, and 7% (n=50,601) Black. Based on geocoded US Census data, 57% (n=409,004) of the cohort had higher household incomes (≥$70,000) and 72% (n=517,624) had some college or higher education. Most of the cohort had commercial insurance (87%, n=561,620), although 3% (n=19,834) had Medicaid. Patients had a median of 21 months (IQR=39) of follow-up data and up to 15 alcohol screenings after entering the registry (Table 2). About 48% (n=346,408) of the cohort had a chronic medical condition, 18% (n=130,031) had a mental health condition, and 4% (n=30,429) had a drug use disorder diagnosis (Table 3). The most common conditions were hypertension (21%, n=152,928), hyperlipidemia (19%, n=134,705), nicotine use disorder (12%, n=86,540), mood disorder (11%, n=82,059), anxiety disorder (11%, n=76,444), and gastroesophageal reflux (10%, n=71,159).

Table 2. Characteristics of patients meeting eligibility criteria for the Kaiser Permanente Northern California Adult Alcohol Registry between 6/1/2013 and 5/31/2019 (N=723,604).

Characteristic	Value
Sex, n (%)[a]	
Male	432,847 (59.8)
Female	290,755 (40.2)
Other/Unknown	2 (<0.1)
Gender, n (%)[a]	
Male	432,614 (59.8)
Female	290,302 (40.1)
Transgender male	217 (<0.1)
Transgender female	241 (<0.1)
Non-binary	229 (<0.1)
Other/Unknown	1 (<0.1)
Age in years, median (IQR)	41.0 (27.0)
Age group (years), n (%)[a]	
18-34	279,276 (38.6)
35-49	187,072 (25.9)
50-64	156,250 (21.6)
≥65	101,006 (14.0)
Race/ethnicity, n (%)[a]	
White	427,606 (59.1)
Asian/Native Hawaiian/Pacific Islander	76,197 (10.5)
Black	50,601 (7.0)
Latino/Hispanic	138,925 (19.2)
Native American	7,015 (1.0)
Other/Unknown	23,260 (3.2)
Household income (US$)[b], n (%)[a]	
0-19,999	5,694 (0.8)
20,000-34,999	38,534 (5.3)
35,000-69,999	264,638 (36.6)
≥70,000	409,004 (56.5)
Unknown	5,734 (0.8)
Education[c], n (%)[a]	
Less than high school	32,446 (4.5)
High school graduate	171,132 (23.6)
Some college or higher	517,624 (71.5)
Unknown	2,402 (0.3)
Smoking status, n (%)[a]	
Never or former	552,618 (76.4)
Current	115,557 (16.0)
Unknown	55,429 (7.7)
Charlson comorbidity score, n (%)[a]	

Characteristic	Value
0	614,422 (84.9)
1	64,420 (8.9)
≥2	44,762 (6.2)
Type of insurance, n (%)[a]	
None	30,033 (4.2)
Medicaid	19,834 (2.7)
Medicare	105,393 (14.6)
Commercial	561,620 (77.6)
Other	6,724 (0.9)
Enrolled via California Affordable Care Act exchange, n (%)[a]	44,110 (6.1)
Months of follow-up data in the registry, median (IQR)	21.0 (39.0)
Number of alcohol screenings, minimum-maximum	0-15

[a]Percentages may not add up to 100% due to rounding error.

[b]Median household income from geocoded census blocks to patients' residential addresses was used as a proxy of individual-level data.

[c]The proportion of individuals within a census block with a level of education was used to estimate each patient's education level.

Table 3. Diagnoses[a] of patients in the Kaiser Permanente Northern California Adult Alcohol Registry (N=723,604).

Condition	Value, n (%)
Chronic medical conditions	
Any chronic medical condition	346,408 (47.9)
Arthritis and other rheumatic conditions	70,371 (9.7)
Asthma	65,073 (9.0)
Atherosclerosis	12,751 (1.8)
Atrial fibrillation	49,141 (6.8)
Cerebrovascular disease	14,920 (2.1)
Chronic kidney disease	23,253 (3.2)
Chronic liver disease	21,363 (3.0)
Chronic obstructive pulmonary disease	21,953 (3.0)
Chronic pain	41,089 (5.7)
Coronary disease	20,644 (2.9)
Dementia	2,143 (0.3)
Diabetes	45,988 (6.4)
Epilepsy	5,050 (0.7)
Gastroesophageal reflux	71,159 (9.8)
Heart failure	8,342 (1.2)
HIV	2,424 (0.3)
Hyperlipidemia	134,705 (18.6)
Hypertension	152,928 (21.1)
Migraine	23,600 (3.3)
Osteoarthritis	66,800 (9.2)
Osteoporosis and osteopenia	18,626 (2.6)
Parkinson’s disease	713 (0.1)
Peptic ulcer	3,074 (0.4)
Rheumatoid arthritis	3,179 (0.4)
Mental health conditions	
Any mental health condition	130,031 (18.0)
Anxiety disorder	76,444 (10.6)
Obsessive-compulsive disorder	1,700 (0.2)
Panic disorder	7,823 (1.1)
Posttraumatic stress disorder	5,312 (0.7)
Eating disorder	924 (0.1)
Anorexia nervosa	276 (<0.1)
Bulimia nervosa	699 (0.1)
Mood disorder	82,059 (11.3)
Bipolar disorder	9,162 (1.3)
Depression	75,445 (10.4)
Other mood disorder	842 (0.1)
Pervasive developmental disorder	221 (<0.1)
Psychoses	6,016 (0.8)
Schizoaffective disorder	1,427 (0.2)

Condition	Value, n (%)
Schizophrenia	1,534 (0.2)
Other psychoses	4,555 (0.6)
Trauma- and stressor-related disorders	12,158 (1.7)
Substance use disorder	
Nicotine use disorder	86,540 (12.0)
Any drug use disorder	30,429 (4.2)
Cannabis	15,175 (2.1)
Cocaine	4,980 (0.7)
Opioid	5,934 (0.8)
Other drugs	10,418 (1.4)
Stimulants	7,293 (1.0)

[a]Diagnoses were identified using ICD codes given at encounters in the year before the patient's eligibility date for the registry (ie, index date).

Discussion

In an integrated health system, we identified a large, population-based cohort of adult patients with unhealthy alcohol use, an AUD, or an alcohol-related health problem that had about 2 years of follow-up time. The KPNC Adult Alcohol Registry can evaluate the full course of alcohol problems, longitudinally and comprehensively, including early identification, initiation and engagement in treatment (including psychiatry, addiction medicine, and pharmacotherapy), and long-term outcomes (eg, drinking, physical and mental well-being), which are critical to understanding recovery. The prevalence of unhealthy alcohol use was 13%, which falls within the range reported by prior studies of the general US population (6%-28%) [16,17]. However, the prevalence of AUD diagnoses in our population (3%) was lower than the 2012-2013 prevalence of Diagnostic and Statistical Manual of Mental Disorders-5 (DSM-5) AUD (13.9% [32]) and DSM-IV AUD (12.6%, [18]) estimated from surveys of the general US population, which might be because diagnoses in health systems depend on clinician assessment and diagnosis during utilization of health care services. Only about 7.6% of individuals with AUD in the general US population seek treatment [34] Additionally, these were crude estimates of prevalence over 6 years and not standardized rates, which a future study could evaluate.

Similar to other studies using population-based survey data that indicated a higher prevalence of unhealthy drinking and AUDs in younger males [18,32], our cohort included more males than females, and younger patients (18-34 years) compared to other age groups. The registry cohort was ethnically diverse, but less representative of lower socioeconomic statuses than samples based on the general US population [18,35]. The cohort included patients with a range of mental health conditions and other substance use disorders, enabling future studies to evaluate the treatment and long-term measurable outcomes in these clinically relevant subgroups.

This EHR-based registry provides a strong foundation for robust research examining the development of alcohol problems and recovery from them. In contrast to national population-based surveys such as the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) [36] and clinical trials such as Project MATCH [37] and COMBINE [38] that collect data from participants at study visits (ie, primary data), the registry takes advantage of data that is collected during health care delivery (ie, secondary data). Primary data collection can be costly for both researchers and participants, especially in large populations, while the use of secondary data can be a cost-effective way to achieve similar research goals. Costs of an EHR-based registry include the initial investment to build it and those related to maintaining it over time, which are less than what a primary research study with equivalent sample size and time points would cost. In many ways, EHR data in KPNC are similar to that in the Veterans Health Administration (VA), the nation's most extensive integrated health care system, which implemented alcohol screening in 2004 [39]; however, our registry cohort of KPNC members is more generalizable to the insured US population since the VA samples are predominantly male, white, and older [40,41].

Additionally, our registry data are longitudinal, spanning over 6 years as of May 31, 2019, and the registry can be continually refreshed with new data extracted from the EHR, including adding new cases and more time points for existing cases. Some current alcohol research studies utilize longitudinal data (eg, NESARC, Project MATCH), but many are repeated cross-sectional studies with different samples (eg, National Health and Nutrition Examination Survey [42], National Health Interview Survey). The registry data are also comprehensive, capturing not only a variety of diagnoses and lab tests that can be used to measure physical functioning, but also health service utilization, insurance factors, and patient-reported outcomes, including alcohol use levels.

We included gender minorities in our registry, given recent research demonstrating a high prevalence of unhealthy drinking in this population [43]. Additionally, the NIAAA has recognized that transgender communities are relevant subpopulations to consider for addressing health disparities [44]. However, there are no general guidelines for how gender minorities should be screened and which risk thresholds to use. For purposes of this registry, we used the patient's EHR-assigned sex, and when

applicable, their response to sex assigned at birth to determine unhealthy alcohol use. We present this approach to be transparent in how sex and gender were operationalized in our registry with the hope of strengthening future research in this area [43].

Limitations

EHR-based registries enable observational studies of "real-world" settings (eg, comparative effectiveness research), an alternative to randomized controlled trials, which may not be feasible; however, using secondary data for research has limitations, including the omission of essential variables and potential for bias (eg, selection bias, information bias, confounding). For example, clinicians in addiction medicine and psychiatry assess AUD symptoms based on the DSM-5, but detailed data are not entered in the EHR. Instead, clinicians record ICD codes to indicate AUD diagnoses, which we use in the registry. While ICD codes for AUDs are not given lightly in other departments, they do occur, and it is not clear what guidelines are used. Therefore, a future validation study of alcohol screening results and the use of these ICD codes is warranted. We also do not have direct measures of individual socioeconomic status (eg, income, education), which are important factors associated with unhealthy alcohol use [18], or social functioning (eg, the Psychosocial Functioning Inventory [45]), an important recovery outcome [46]. Though not only an issue with secondary data analysis, missing data can create bias in a study if it is not missing completely at random; therefore, future studies utilizing the registry data should check for missingness and apply proper statistical methods to address issues as needed [47]. Reliance on accurate reporting of alcohol use and other measures is also a concern; however, it is not a unique problem of EHR data and shared by other studies that collect self-report data. While novel statistical methodologies can be applied to deal with issues of confounding [48] (eg, the counterfactual framework [49]), temporality may remain an issue. Measures of alcohol use and diagnoses are recorded in the EHR when patients seek care rather than when alcohol-related issues emerge, similar to other disease-based registries that rely on data collected during care (eg, diagnostic tests for cancer) and survey-based studies that gather data on past-year or lifetime alcohol problems without specific dates. Sex and gender variables in the EHR can change and are not collected longitudinally, so their values in the registry reflect what was present at the time of the data extraction rather than historical values, for example, at the time of alcohol screening. We are also not certain which variables are used to determine the appropriate screening questions and risk thresholds (especially for gender minorities), which a future study could evaluate. Therefore, some alcohol screening results may have been misclassified in the registry, affecting eligibility; however, we expect this to have a minimal impact on future studies.

Future Directions

While we included only core data elements that were necessary to address our research aims, the registry could be extended to include other types of data, including provider information, family members of patients with alcohol problems, and medications prescribed off-label to treat AUD (eg, gabapentin [50]). Other health systems in the HCSRN with harmonized VDW data [22] may also want to create their own registry of alcohol problems, enabling the potential for multi-site studies [51-53]. The registry's utility may also extend beyond research to clinical or administrative purposes, for example, to manage care or evaluate performance, which would require additional support from KPNC organizational stakeholders and institutional review boards to protect patient privacy and confidentiality.

Conclusions

We demonstrate that EHR-based data collected during routine clinical care within an integrated health care system can be leveraged to develop a registry of patients with alcohol problems that is flexible and can be easily refreshed and extended. The registry can be used to address critical public health questions related to the full spectrum and course of alcohol problems, which will complement other methods used in alcohol research. Future analyses will aim to provide insight on how to strengthen efforts in the prevention of alcohol-related disability and mortality and improve patient-centered health care delivery. We hope that other researchers and health systems interested in assembling a similar registry can take advantage of the time we invested in developing this protocol.

Acknowledgments

This project was funded by contracts (#HHSN275201800625P and #75N94019P00907) and a grant (R01AA025902) from the National Institute on Alcohol Abuse and Alcoholism (NIAAA). We gratefully acknowledge Dr Raye Litten and Dr Daniel Falk at the NIAAA for their expertise in alcohol research, and Yun Lu at the Kaiser Permanente Northern California Division of Research for assistance in extracting data used in this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed descriptions of data elements in the Kaiser Permanente Northern California Adult Alcohol Registry.
[DOCX File , 23 KB - medinform_v8i7e19081_app1.docx]

Multimedia Appendix 2
Additional diagnoses tracked among patients in the Kaiser Permanente Northern California Adult Alcohol Registry.
[DOCX File , 14 KB - medinform_v8i7e19081_app2.docx]

References

1. Schmittdiel J, Bodenheimer T, Solomon NA, Gillies RR, Shortell SM. Brief report: The prevalence and use of chronic disease registries in physician organizations. A national survey. J Gen Intern Med 2005 Sep;20(9):855-858 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.0171.x] [Medline: 16117756]
2. McEvoy P, Laxade S. Patient registries: a central component of the chronic care model. Br J Community Nurs 2008 Mar;13(3):127-8, 130. [doi: 10.12968/bjcn.2008.13.3.28677] [Medline: 18557574]
3. Feller DJ, Lor M, Zucker J, Yin MT, Olender S, Ferris DC, et al. An investigation of the information technology needs associated with delivering chronic disease care to large clinical populations. Int J Med Inform 2020 Feb 13;137:104099. [doi: 10.1016/j.ijmedinf.2020.104099] [Medline: 32088558]
4. Gliklich RE, Dreyer NA, Leavy MB, editors. Section I. Creating Registries. In: Registries for Evaluating Patient Outcomes: A User's Guide. 3rd ed. Rockville, MD: Agency for Healthcare Research and Quality (US); Apr 2014.
5. Global status report on alcohol and health 2018. Geneva: World Health Organization; 2018. URL: https://apps.who.int/iris/bitstream/handle/10665/274603/9789241565639-eng.pdf?ua=1 [accessed 2019-09-10]
6. White AM, Castle IP, Hingson RW, Powell PA. Using Death Certificates to Explore Changes in Alcohol-Related Mortality in the United States, 1999 to 2017. Alcohol Clin Exp Res 2020 Jan;44(1):178-187. [doi: 10.1111/acer.14239] [Medline: 31912524]
7. Spillane S, Shiels MS, Best AF, Haozous EA, Withrow DR, Chen Y, et al. Trends in Alcohol-Induced Deaths in the United States, 2000-2016. JAMA Netw Open 2020 Feb 05;3(2):e1921451 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.21451] [Medline: 32083687]
8. Samokhvalov AV, Rehm J, Roerecke M. Alcohol Consumption as a Risk Factor for Acute and Chronic Pancreatitis: A Systematic Review and a Series of Meta-analyses. EBioMedicine 2015 Dec;2(12):1996-2002 [FREE Full text] [doi: 10.1016/j.ebiom.2015.11.023] [Medline: 26844279]
9. Ricci C, Wood A, Muller D, Gunter MJ, Agudo A, Boeing H, et al. Alcohol intake in relation to non-fatal and fatal coronary heart disease and stroke: EPIC-CVD case-cohort study. BMJ 2018 May 29;361:k934 [FREE Full text] [doi: 10.1136/bmj.k934] [Medline: 29844013]
10. Shield KD, Soerjomataram I, Rehm J. Alcohol Use and Breast Cancer: A Critical Review. Alcohol Clin Exp Res 2016 Jun;40(6):1166-1181. [doi: 10.1111/acer.13071] [Medline: 27130687]
11. Shield KD, Parry C, Rehm J. Chronic diseases and conditions related to alcohol use. Alcohol Res 2013;35(2):155-173 [FREE Full text] [Medline: 24881324]
12. Thomas RM, Francis Gerstel PA, Williams EC, Sun H, Bryson CL, Au DH, et al. Association between alcohol screening scores and diabetic self-care behaviors. Fam Med 2012 Sep;44(8):555-563 [FREE Full text] [Medline: 22930120]
13. Worthington J, Fava M, Agustin C, Alpert J, Nierenberg AA, Pava JA, et al. Consumption of alcohol, nicotine, and caffeine among depressed outpatients. Relationship with response to treatment. Psychosomatics 1996;37(6):518-522. [doi: 10.1016/S0033-3182(96)71515-3] [Medline: 8942202]
14. Sullivan LE, Fiellin DA, O'Connor PG. The prevalence and impact of alcohol problems in major depression: a systematic review. Am J Med 2005 Apr;118(4):330-341. [doi: 10.1016/j.amjmed.2005.01.007] [Medline: 15808128]
15. Bahorik AL, Leibowitz A, Sterling SA, Travis A, Weisner C, Satre DD. The role of hazardous drinking reductions in predicting depression and anxiety symptom improvement among psychiatry patients: A longitudinal study. J Affect Disord 2016 Dec;206:169-173 [FREE Full text] [doi: 10.1016/j.jad.2016.07.039] [Medline: 27475887]
16. Helping patients who drink too much: a clinician's guide. Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 2005. URL: https://pubs.niaaa.nih.gov/publications/practitioner/cliniciansguide2005/guide.pdf [accessed 2018-03-13]
17. Azagba S, Shan L, Latham K, Manzione L. Trends in Binge and Heavy Drinking among Adults in the United States, 2011-2017. Subst Use Misuse 2020 Jan 30:1-8. [doi: 10.1080/10826084.2020.1717538] [Medline: 31999198]
18. Grant BF, Chou SP, Saha TD, Pickering RP, Kerridge BT, Ruan WJ, et al. Prevalence of 12-Month Alcohol Use, High-Risk Drinking, and DSM-IV Alcohol Use Disorder in the United States, 2001-2002 to 2012-2013: Results From the National Epidemiologic Survey on Alcohol and Related Conditions. JAMA Psychiatry 2017 Sep 01;74(9):911-923 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.2161] [Medline: 28793133]
19. Mertens JR, Chi FW, Weisner CM, Satre DD, Ross TB, Allen S, et al. Physician versus non-physician delivery of alcohol screening, brief intervention and referral to treatment in adult primary care: the ADVISe cluster randomized controlled implementation trial. Addict Sci Clin Pract 2015 Nov 19;10:26 [FREE Full text] [doi: 10.1186/s13722-015-0047-0] [Medline: 26585638]
20. Gordon N. Similarity of the adult Kaiser Permanente membership in Northern California to the insured and general population in Northern Californiatatistics from the 2011 California Health Interview Survey. 2015. URL: https://divisionofresearch.kaiserpermanente.org/projects/memberhealthsurvey/SiteCollectionDocuments/chis_non_kp_2011.pdf [accessed 2019-03-15]

21. Chi FW, Satre DD, Weisner C. Chemical dependency patients with cooccurring psychiatric diagnoses: service patterns and 1-year outcomes. Alcohol Clin Exp Res 2006 May;30(5):851-859. [doi: 10.1111/j.1530-0277.2006.00100.x] [Medline: 16634854]
22. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration. EGEMS (Wash DC) 2014;2(1):1049. [doi: 10.13063/2327-9214.1049] [Medline: 25848584]
23. Viviani L, Zolin A, Mehta A, Olesen HV. The European Cystic Fibrosis Society Patient Registry: valuable lessons learned on how to sustain a disease registry. Orphanet J Rare Dis 2014 Jun 07;9:81 [FREE Full text] [doi: 10.1186/1750-1172-9-81] [Medline: 24908055]
24. Gitt AK, Bueno H, Danchin N, Fox K, Hochadel M, Kearney P, et al. The role of cardiac registries in evidence-based medicine. Eur Heart J 2010 Mar;31(5):525-529. [doi: 10.1093/eurheartj/ehp596] [Medline: 20093258]
25. Silverberg MJ, Chao C, Leyden WA, Xu L, Horberg MA, Klein D, et al. HIV infection, immunodeficiency, viral replication, and the risk of cancer. Cancer Epidemiol Biomarkers Prev 2011 Dec;20(12):2551-2559 [FREE Full text] [doi: 10.1158/1055-9965.EPI-11-0777] [Medline: 22109347]
26. Karter AJ, Ackerson LM, Darbinian JA, D'Agostino RB, Ferrara A, Liu J, et al. Self-monitoring of blood glucose levels and glycemic control: the Northern California Kaiser Permanente Diabetes registry. Am J Med 2001 Jul;111(1):1-9. [doi: 10.1016/s0002-9343(01)00742-2] [Medline: 11448654]
27. Oehrli M, Quesenberry C, Leyden W. Annual report on trends, incidence, and outcomes: Northern California Cancer Registry at the Division of Research. Oakland, CA: Kaiser Permanente Northern California; 2018.
28. Ray GT, Bahorik AL, VanVeldhuisen PC, Weisner CM, Rubinstein AL, Campbell CI. Prescription opioid registry protocol in an integrated health system. Am J Manag Care 2017 May 01;23(5):e146-e155 [FREE Full text] [Medline: 28810131]
29. Alcohol and Public Health: Alcohol-Related Disease Impact. Alcohol-related ICD codes. Atlanta, GA: Centers for Disease Control and Prevention; 2019. URL: https://nccd.cdc.gov/DPH_ARDI/Info/ICDCodes.aspx [accessed 2019-01-12]
30. Weisner C, Campbell CI, Altschuler A, Yarborough BJH, Lapham GT, Binswanger IA, et al. Factors associated with Healthcare Effectiveness Data and Information Set (HEDIS) alcohol and other drug measure performance in 2014-2015. Subst Abus 2019;40(3):318-327 [FREE Full text] [doi: 10.1080/08897077.2018.1545728] [Medline: 30676915]
31. Rossiter K. What are census blocks?.: United States Census Bureau; 2011. URL: https://www.census.gov/newsroom/blogs/random-samplings/2011/07/what-are-census-blocks.html [accessed 2020-03-23]
32. Grant BF, Goldstein RB, Saha TD, Chou SP, Jung J, Zhang H, et al. Epidemiology of DSM-5 Alcohol Use Disorder: Results From the National Epidemiologic Survey on Alcohol and Related Conditions III. JAMA Psychiatry 2015 Aug;72(8):757-766 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.0584] [Medline: 26039070]
33. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992 Jun;45(6):613-619. [Medline: 1607900]
34. Olfson M, Blanco C, Wall MM, Liu S, Grant BF. Treatment of Common Mental Disorders in the United States: Results From the National Epidemiologic Survey on Alcohol and Related Conditions-III. J Clin Psychiatry 2019 May 28;80(3) [FREE Full text] [doi: 10.4088/JCP.18m12532] [Medline: 31141319]
35. Chavez LJ, Bradley K, Tefft N, Liu C, Hebert P, Devine B. Preference weights for the spectrum of alcohol use in the U.S. Population. Drug Alcohol Depend 2016 Apr 01;161:206-213. [doi: 10.1016/j.drugalcdep.2016.02.004] [Medline: 26900145]
36. National Epidemiologic Survey on Alcohol and Related Conditions-III (NESARC-III). Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 2015. URL: https://www.niaaa.nih.gov/research/nesarc-iii [accessed 2019-09-27]
37. Project MATCH Research Group. Project MATCH (Matching Alcoholism Treatment to Client Heterogeneity): rationale and methods for a multisite clinical trial matching patients to alcoholism treatment. Alcohol Clin Exp Res 1993 Dec;17(6):1130-1145. [doi: 10.1111/j.1530-0277.1993.tb05219.x] [Medline: 8116822]
38. COMBINE Study Research Group. Testing combined pharmacotherapies and behavioral interventions in alcohol dependence: rationale and methods. Alcohol Clin Exp Res 2003 Jul;27(7):1107-1122. [doi: 10.1097/00000374-200307000-00011] [Medline: 12878917]
39. Bradley KA, Williams EC, Achtmeyer CE, Volpp B, Collins BJ, Kivlahan DR. Implementation of evidence-based alcohol screening in the Veterans Health Administration. Am J Manag Care 2006 Oct;12(10):597-606 [FREE Full text] [Medline: 17026414]
40. Chavez LJ, Williams EC, Lapham GT, Rubinsky AD, Kivlahan DR, Bradley KA. Changes in Patient-Reported Alcohol-Related Advice Following Veterans Health Administration Implementation of Brief Alcohol Interventions. J Stud Alcohol Drugs 2016 May;77(3):500-508 [FREE Full text] [doi: 10.15288/jsad.2016.77.500] [Medline: 27172583]
41. Kalpakci A, Sofuoglu M, Petrakis I, Rosenheck RA. Gender differences among Veterans with alcohol use disorder nationally in the Veterans Health Administration. J Addict Dis 2018;37(3-4):185-194. [doi: 10.1080/10550887.2019.1653739] [Medline: 31429377]
42. National Health and Nutrition Examination Survey. National Center for Health Statistics. Atlanta, GA: Centers for Disease Control and Prevention; 2019. URL: https://www.cdc.gov/nchs/nhanes/index.htm [accessed 2019-09-26]

43. Gilbert PA, Pass LE, Keuroghlian AS, Greenfield TK, Reisner SL. Alcohol research with transgender populations: A systematic review and recommendations to strengthen future studies. Drug Alcohol Depend 2018 May 01;186:138-146 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.01.016] [Medline: 29571076]
44. National Institute on Alcohol Abuse and Alcoholism Strategic Plan 2017-2021. Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 2017. URL: https://www.niaaa.nih.gov/sites/default/files/StrategicPlan_NIAAA_optimized_2017-2020.pdf [accessed 2019-09-25]
45. Feragne MA, Longabaugh R, Stevenson JF. The psychosocial functioning inventory. Eval Health Prof 1983 Mar;6(1):25-48. [doi: 10.1177/016327878300600102] [Medline: 10259949]
46. Witkiewitz K, Kirouac M, Roos CR, Wilson AD, Hallgren KA, Bravo AJ, et al. Abstinence and low risk drinking during treatment: Association with psychosocial functioning, alcohol use, and alcohol problems 3 years following treatment. Psychol Addict Behav 2018 Sep;32(6):639-646 [FREE Full text] [doi: 10.1037/adb0000381] [Medline: 30160499]
47. Rubin DB. Inference and Missing Data. Biometrika 1976 Dec;63(3):581. [doi: 10.2307/2335739]
48. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002 Jan 19;359(9302):248-252. [doi: 10.1016/S0140-6736(02)07451-2] [Medline: 11812579]
49. Bours MJL. A nontechnical explanation of the counterfactual definition of confounding. J Clin Epidemiol 2020 Feb 14;121:91-100. [doi: 10.1016/j.jclinepi.2020.01.021] [Medline: 32068101]
50. Anton RF, Latham P, Voronin K, Book S, Hoffman M, Prisciandaro J, et al. Efficacy of Gabapentin for the Treatment of Alcohol Use Disorder in Patients With Alcohol Withdrawal Symptoms: A Randomized Clinical Trial. JAMA Intern Med 2020 Mar 09. [doi: 10.1001/jamainternmed.2020.0249] [Medline: 32150232]
51. Binswanger IA, Carroll NM, Ahmedani BK, Campbell CI, Haller IV, Hechter RC, et al. The association between medical comorbidity and Healthcare Effectiveness Data and Information Set (HEDIS) measures of treatment initiation and engagement for alcohol and other drug use disorders. Subst Abus 2019;40(3):292-301 [FREE Full text] [doi: 10.1080/08897077.2018.1545726] [Medline: 30676892]
52. Health Care Systems Research Network. The Virtual Data Warehouse (VDW) and how to use it. URL: http://www.hcsrn.org/en/Resources/VDW/VDWScientists/The+VDW+and+How+to+Use+It.pdf
53. Yarborough BJH, Ahmedani BK, Boggs JM, Beck A, Coleman KJ, Sterling S, et al. Challenges of Population-based Measurement of Suicide Prevention Activities Across Multiple Health Systems. EGEMS (Wash DC) 2019 Apr 12;7(1):13 [FREE Full text] [doi: 10.5334/egems.277] [Medline: 30993146]

Abbreviations

AUD: alcohol use disorder
EHR: electronic health record
HCSRN: Health Care Systems Research Network
ICD: International Classification of Diseases
ICD-9: International Classification of Diseases, 9th Revision, Clinical Modification
ICD-10: International Classification of Diseases, 10th Revision, Clinical Modification
KP: Kaiser Permanente
KPNC: Kaiser Permanente Northern California
NESARC: National Epidemiologic Survey on Alcohol and Related Conditions
NIAAA: National Institute on Alcohol Abuse and Alcoholism
VDW: Virtual Data Warehouse

Edited by C Lovis; submitted 02.04.20; peer-reviewed by K Hallgren, K Phillips; comments to author 26.04.20; revised version received 08.05.20; accepted 11.05.20; published 22.07.20.

Please cite as:
Palzes VA, Weisner C, Chi FW, Kline-Simon AH, Satre DD, Hirschtritt ME, Ghadiali M, Sterling S
The Kaiser Permanente Northern California Adult Alcohol Registry, an Electronic Health Records-Based Registry of Patients With Alcohol Problems: Development and Implementation
JMIR Med Inform 2020;8(7):e19081
URL: http://medinform.jmir.org/2020/7/e19081/
doi:10.2196/19081
PMID:32706676

Original Paper

A Multiview Model for Detecting the Inappropriate Use of Prescription Medication: Machine Learning Approach

Lin Zhuo[1,2*], PhD; Yinchu Cheng[3*], PhD; Shaoqin Liu[4], BA; Yu Yang[5], PhD; Shuang Tang[4], BA; Jiancun Zhen[6], MS; Junfeng Zhao[4], PhD; Siyan Zhan[1,2], MD, PhD

[1]Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing, China
[2]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China
[3]Department of Pharmacy, Peking University Third Hospital, Beijing, China
[4]School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[5]Center for Data Science in Medicine and Health, Peking University, Beijing, China
[6]Department of Pharmacy, Ji Shui Tan Hospital and Fourth Medical College of Peking University, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Siyan Zhan, MD, PhD
Research Center of Clinical Epidemiology
Peking University Third Hospital
49 North Garden Rd, Haidian District
Beijing, 100191
China
Phone: 86 1082805162
Email: siyan-zhan@bjmu.edu.cn

Abstract

Background: The inappropriate use of prescription medication has recently garnered worldwide attention, but most national policies do not effectively provide for early detection or timely intervention.

Objective: This study aimed to develop and assess the validity of a model that can detect the inappropriate use of prescription medication. This effort combines a multiview and topic matching method. The study also assessed the validity of this approach.

Methods: A multiview extension of the latent Dirichlet allocation algorithm for topic modeling was chosen to generate diagnosis-medication topics, with data obtained from the Chinese Monitoring Network for Rational Use of Drugs (CMNRUD) database. Topic mapping allowed for calculating the degree to which diagnoses and medications were similarly distributed and, by setting a threshold, for identifying prescription misuse. The Beijing Regional Prescription Review Database (BRPRD) database was used as the gold standard to assess the model's validity. We also conducted a sensitivity analysis using random samples of validated prescriptions and evaluated the model's performance.

Results: A total of 44 million prescriptions were used to generate topics using the diagnoses and medications from the CMNRUD database. A random sample (15,000 prescriptions) from the BRPRD was used for validation, and it was found that the model had a sensitivity of 81.8%, specificity of 47.4%, positive-predictive value of 14.5%, and negative-predictive value of 96.0%. The model showed superior stability under different sampling proportions.

Conclusions: A method that combines multiview topic modeling and topic matching can detect the inappropriate use of prescription medication. This model, which has mediocre specificity and moderate sensitivity, can be used as a primary screening tool and will likely complement and improve the process of manually reviewing prescriptions.

(JMIR Med Inform 2020;8(7):e16312) doi:10.2196/16312

KEYWORDS

inappropriate use of prescription medication; topic model; latent Dirichlet allocation; multiview learning; prescription review

Introduction

It is estimated that more than 50% of medicines are inappropriately prescribed, dispensed, or sold, which represents a universal challenge for medical practice [1]. Furthermore, in developing countries, the treatment of about 60% to 70% of patients in primary care does not meet standard treatment guidelines [2]. This inappropriate use of prescription is wasteful and costly, and can increase the risk of adverse drug reactions [3]. Finally, the overuse of antimicrobial and antibiotic injections may result in certain pathogens developing antibiotic resistance [4].

The excessive use of antibiotics is common in China, as is the injection of traditional Chinese medicines [5-7]. Antibiotics are present in 50% of prescriptions and injectable medicines in 30%, exceeding the World Health Organization's standard treatment guidelines [7]. The Chinese government released the *Management Practices of Hospital Prescription Comment (Trial)* in 2010 to assess compliance with rational criteria for using prescription drugs [8]. This document requires that each hospital assign trained pharmacists monthly to review a minimum of 100 randomly sampled prescriptions. However, these reviews are currently associated with limited coverage, high omission rates, a lack of representativeness, and supervisory lag. All of this points to the urgency of improving the review process, especially when Chinese hospitals are witnessing a continuous daily increase in outpatient prescriptions [9,10].

A few knowledge-based approaches have been implemented in the health care information systems (HISs) of Chinese hospitals to screen the appropriateness of prescriptions [11,12]. Prior knowledge, including treatment guidelines, formularies, package inserts, expert knowledge, and published literature, indicates that these systems are generally working well. However, they are time-consuming and costly to establish and maintain [13]. Furthermore, timely updates to these systems are challenging because of the continuous availability of both new drugs and new research.

Currently, both supervised and unsupervised data-driven methodologies, whether as alternatives or supplements to the systems listed above, are being used to identify outliers and detect inappropriate prescriptions. Such approaches remain constrained, however, by the difficulty associated with using supervised methods to obtain high-quality labeled sample data [14,15]. Other limiting factors include defining outliers and considering their association rules, which effectively account for the relationships between features [16]. Usually, diagnosis and medication are closely and consistently related to the clinical condition of the patient. In the absence of this consistency, prescriptions are more likely to be inappropriate (or anomalous). Contextual anomaly detection is one approach to capture the relationship between features (eg, between medication and diagnosis) and to detect exceptions caused by feature mismatch [17,18]. Nonetheless, this does not work well with prescription data or similar information in a high-dimensional sparse space [19].

By contrast, a topic modeling method, the latent Dirichlet allocation (LDA) method [20], has been proven to be useful in dimensional reduction when mining patient records [21]. Here, a "topic" is defined as a collection of semantically related terms that appear frequently and relate to a common subject [22]. LDA, a probabilistic statistical model with the assumption that topic distributions are drawn from their prior distributions, can be used to describe the composition of high-dimensional unstructured text and to capture clusters of words that reveal critical concepts [21,23]. Beginning with its appearance in the biomedical domain, LDA has been used in mining clinical pathway patterns [24-26], image processing [27-29], risk stratification [30], and bioinformatics [31-33]. One drawback of LDA is that it cannot simultaneously consider both diagnosis and medication. We therefore adopted a multiview [23,34-36] concept that enhances the topic modeling capacity of LDA and coordinated it with anomaly detection techniques to build a multiview LDA model (MV-LDA). This model, which was tested in our previous simulation study, had a greater area under the precision-recall curve [37] than the two traditional methods (point anomaly detection and contextual anomaly detection) and had better suitability for high-dimensional sparse data.

Methods

Data Sources

One subsystem of the Chinese Monitoring Network for the Rational Use of Drugs (CMNRUD) was the data source for model development. The CMNRUD was launched by the Chinese Ministry of Health in 2010, and it covers over 86% (30/35) of the provinces in China [38], including 60% of the nation's tertiary hospitals (955 hospitals) and 6% of its secondary hospitals (375 hospitals) [39]. Each monitoring hospital must upload encrypted data every month. The system organizes the prescriptions in a stipulated uniform structure, and thereafter, some of the cleaned data are checked by data management professionals. Since 2013, the system has been performing automatic uploading, preliminary cleaning, recoding, and verification.

The CMNRUD consists of the following four monitoring subsystems: outpatient prescriptions, clinical drug use, medical damage, and critical disease. Anonymized data from outpatient prescriptions (from October to December 2016) were used to build the model for detecting prescription misuse. These data include demographic, diagnostic, and drug-related information. Diagnostic information includes the patient ID, diagnosis date, diagnosis description, and diagnostic code (10th revision of the International Classification of Diseases, ICD-10), which are directly related to the purpose of the patient's visit or their condition. It sometimes may not include other complications that do not require further treatment. A higher diagnosis ranking was associated with more visit relevance. Available drug-related information (no more than five medications per prescription) includes information such as the prescription date, generic and brand names, corresponding Anatomical Therapeutic Chemical code, dosage, and administration route. The medications were listed randomly, preventing them from being mapped to the

corresponding diagnoses on a one-to-one basis. The variables taken from the CMNRUD are presented in Table 1.

The data for model validation were randomly selected from the Beijing Regional Prescription Review Database (BRPRD) [40], which was created by the Beijing Municipal Administration of Hospitals in 2010. The BRPRD extracts 1 week of prescriptions every quarter from the HISs of 17 tertiary hospitals and five secondary hospitals in Beijing. A total of 19 hospitals from the BRPRD were included among the 65 CMNRUD monitoring hospitals from Beijing and accounted for 5.4% (19/349) of the hospitals in the entire CMNRUD database. The prescription variables include treatment type, prescription number, prescription date, age, sex, diagnosis, and medication. As part of the standard procedure, a prescription review board of trained clinicians and clinical pharmacists regularly examines the prescriptions individually based on a standardized guideline and then captures inappropriate data in the BRPRD database [8,41].

Table 1. Main variables in the Chinese Monitoring Network for the Rational Use of Drugs outpatient prescription monitoring subsystem.

Information	Variables[a]
Basic information (patients)	Patient ID, treatment card number, sex, age, and age range
Basic information (hospital, department, and doctor)	Hospital name, hospital grade, hospital type, region, department, and doctor ID
Diagnosis	Diagnosis name, and ICD-10[b] by class, suborder, and type
Medication[c]	Prescription ID, prescription type, prescription date, ATC[d] code, drug trade name, drug generic name, specification, quantity, unit, dosage, usage, price, pharmaceutical company, and individual hospital information

[a]The variables indicate the features of the multiview latent Dirichlet allocation model.

[b]ICD-10: 10th revision of the International Classification of Diseases.

[c]Since a generic medicine works the same as its branded version and owing to the limitation of the Anatomical Therapeutic Chemical's lack of codes for traditional Chinese medicine, we used generic names, which are well recorded in the database, to build the topic model.

[d]ATC: Anatomical Therapeutic Chemical.

Study Approval

The Institutional Review Board of Peking University reviewed and approved the study protocol before the study commenced, and it determined that informed consent was not required (reference number: IRB00001052-17003-Exempt).

Study Design

We developed and evaluated an MV-LDA model for detecting the inappropriate use of prescription medications in the following four steps: (1) data preparation, gathering and cleaning data from the CMNRUD and BRPRD database; (2) topic generation, using MV-LDA topic-modeling methods and CMNRUD data to extract associations between diagnoses and medications; (3) inferring and anomaly scoring, using BRPRD data and the topics extracted in step 2 to infer the distribution of each prescription and measuring the degree to which diagnoses and medications show similar distributions (less similarity is associated with more likelihood that the item represents the inappropriate use of prescription medication); (4) model evaluation and sensitivity analysis, evaluating the model for detecting prescription misuse with the results of the BRPRD review.

Step 1: Data Preparation

Prescriptions between October 2016 and December 2016 from CMNRUD were used, and those with missing prescription identifiers or with medication withdrawal were excluded. The patient ID, treatment card number, prescription date, diagnosis name, and generic drug name were chosen to build the topics.

For model evaluation, considering the sensitivity and specificity (71.5% and 68.8%, respectively) of the Apriori algorithm in previous work [42,43], we set both the expected sensitivity and specificity at 80%. By setting a significance level of .05 and an allowable error of 0.05, we needed at least 14,471 prescriptions given 1.7% prevalence [44] of prescription misuse according to equation 1.

Finally, we randomly selected a total of 15,000 prescriptions from 2016 BRPRD data that had already been manually reviewed by experienced pharmacists, with the prescriptions that included the following three variables: prescription ID, diagnosis, and medication.

Step 2: Topic Generation

The study is based on the assumption that the prescription database is mostly composed of regular instances (ie, rational appropriate prescriptions), and a probabilistic model is fitted to all features.

LDA assumes that a set of documents or instances exhibits a specific number of latent independent topics, and then, the given topics generate the terms probabilistically. A graphical representation of the LDA model is given in Figure 1. With specific input records and setting hyperparameters α and β, LDA can detect *K* topics, formally presented as two multinomial distributions (topic-word distribution ϕ and document-topic distribution θ). The LDA model formula is shown in equation 2.

Like the standard LDA, MV-LDA can be represented as a probability pattern, as shown in Figure 2. In the medical domain, we can consider a clinical condition (topic) as a probability distribution over related diagnostic codes, and a patient's diagnostic record can be regarded as a "document" composed of different clinical topics. The same applies to medication. In our study, we used CMNRUD data to generate MV-LDA topics. For prescription *m*, features A and B represent the diagnoses and medications, respectively. Both follow the same generative process mentioned above (ie, they comply with the same topic distribution θ), and then, α and β become the hyperparameters of prescription-topic distribution and topic-diagnosis (or topic-medication) distribution within topics. Moreover, ϕ^A and ϕ^B represent the topic feature distribution of A and B, respectively. In summary, the MV-LDA model is a combination of two separate LDA models (here they are called f^A and f^B) integrated by the common distribution θ. In Figure 2, N^A and N^B are the total numbers of diagnoses and medications with each prescription; this value can only be a discrete integer.

Figure 1. Graphical representation of the latent Dirichlet allocation model. K: number of topics; M: number of documents; N: number of words in each document; x: observed words in the document m; z: topic of nth word in a document m; θ: topic distribution for document m (document-topic distribution); ϕ: topic-word distribution; α: hyperparameter of θ; β: hyperparameter of ϕ.

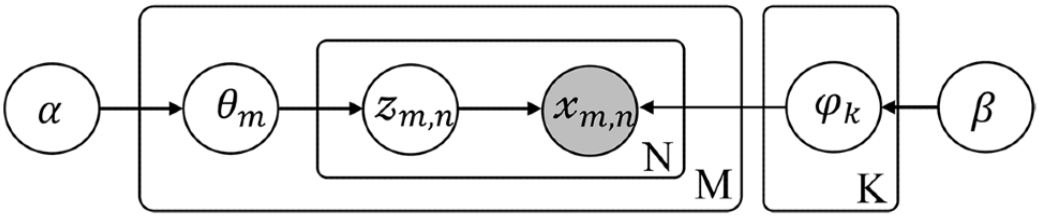
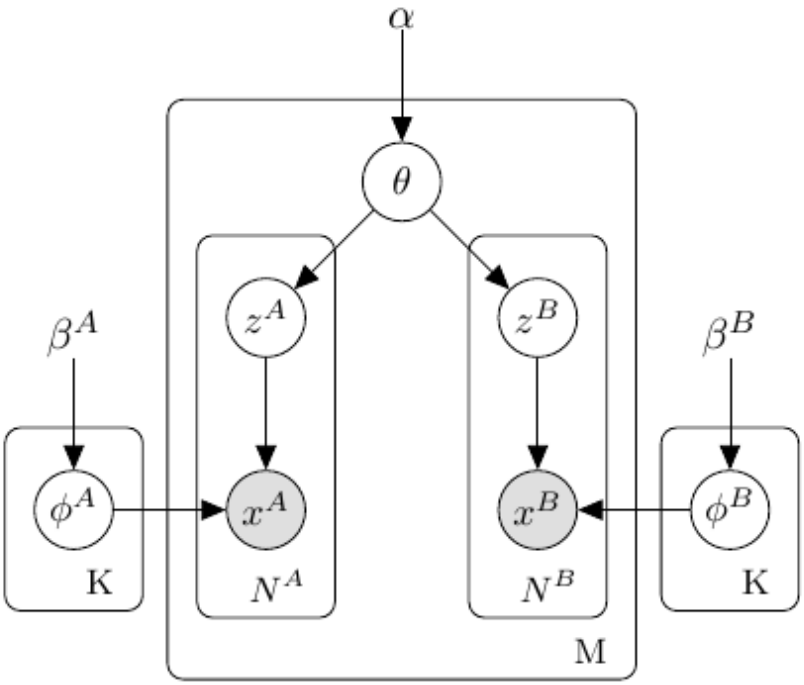


Figure 2. Graphical representation of the multiview latent Dirichlet allocation model. K: number of topics; M: number of prescriptions; N^A: number of diagnoses per prescription; N^B: number of medications per prescription; x^A: diagnosis (type A feature); x^B: medication (type B feature); z^A: topic of x^A; z^B: topic of x^B; ϕ^A: topic-diagnosis distribution; ϕ^B: topic-medication distribution; β^A: hyperparameter of ϕ^A; β^B: hyperparameter of ϕ^B; θ: prescription-topic distribution; α: hyperparameter of θ.



The MV-LDA model generates topics as follows: (1) For each topic, draw features $\phi^A \sim Dirichlet(\beta^A)$ and draw features $\phi^B \sim Dirichlet(\beta^B)$; (2) For each prescription, draw topic proportions $\theta \sim Dirichlet(\alpha)$; for each feature A, draw $z_{m,n} \sim Mult(\theta_m)$ and draw $x_{m,n} \sim Mult(\phi_z^A)$; and for each feature B, repeat the steps for feature A.

We adopted Gibbs sampling to create the model and parameters ϕ and θ. Topics were first randomly assigned to all of the features. Every diagnosis (type A feature) or medication (type B feature) in each prescription that corresponds to a topic is iteratively sampled. The calculation of the conditional probability of x^A is shown in equation 3, with the related notations shown in Table 2. Here, the first factor of equation 3 only accounts for the type A feature topic-diagnosis counts, whereas the second factor calculates the prescription-topic for all features. This formula also applies to type B features.

The features of both types (A and B) were iteratively sampled for each prescription until the model converged. Thereafter, we utilized the result to calculate the parameters ϕ^A and ϕ^B, which are used for the inferring step, and ϕ^A can be calculated as in equation 4. We set eight topic numbers (15, 20, 25, 30, 35, 40, 45, and 50) and built the MV-LDA model according to previous research and a pilot study revealing that LDA had a moderate ability in terms of generating topics for electronic medical records with a topic number of around 30 [45].

Table 2. Notations of the multiview latent Dirichlet allocation model in Gibbs sampling.

Variable	Description
K	Topic number
V^A	Number of diagnoses (type A feature)
	Number of times that x^A is assigned to topic k
	Number of times that any feature A is assigned to topic k
	Number of all features of prescription m (including both A and B) assigned to topic k
	All features in prescription m

Step 3: Inferring and Anomaly Scoring

In this step, a separate dataset was used for inferring 15,000 randomly sampled prescriptions from the BRPRD (2016) that had already been manually reviewed by experienced pharmacists. Each feature in the MV-LDA model can be treated as an independent LDA model and can be inferred separately. To be specific, for the MV-LDA model obtained in the previous learning step, ϕ^A can be used to detect the new prescriptions in question, but this only pertains to estimations of the topic distribution under feature A. The equation for this is as follows:

In this equation, $\phi_{x,k}{}^A$ is the value of the topic distribution under the circumstance of topic k and feature x. Finally, as in the topic generating step (step 2), we inferred the marginal θ based on the Gibbs sampling shown in equation 5, which indicated the proportion of feature A assigned to topic k for each prescription. Additionally,  is calculated under type B features. For each test prescription, both  and  were inferred and used to calculate the anomaly score.

The assumption mentioned in step 2 is that the given order of diagnoses and prescribed medication should show consistency (ie, the values for  and  should be equal or close to each other), and if not, the prescription might be inappropriate. The similarity between  and  was measured using novel topic mapping (TM) methods [46]. TM was performed in the following manner: we allocated topic feature distributions from the MV-LDA model for every diagnosis or medication before matching. First, high probability topics were tagged for each diagnosis. Thereafter, we similarly identified the most probable topics for each medication and added up the total. When a topic was not tagged, it was assigned an anomaly score of 1. Finally, the anomaly scores for each prescription were summed, and different thresholds were used to filter potentially inappropriate prescriptions.

Step 4: Model Evaluation and Sensitivity Analysis

The same prescriptions (15,000 randomly sampled prescriptions from the BRPRD in 2016) were inferred and detected by the MV-LDA model. Multimedia Appendix 1 shows the confusion matrix of the screening test we used. The sensitivity, specificity, positive-predictive value (PPV), negative-predictive value (NPV), and Youden's index were computed from the results to compare the assessments between the model and the experts and to identify the best performance parameter setting of TM. A sensitivity analysis was performed by randomly sampling 90%, 70%, 50%, 30%, and 10% of prescriptions from the evaluation data of the 15,000 prescriptions. The sensitivity, specificity, PPV, NPV, Youden's index, and area under the receiver operating characteristic curve were compared. It should be noted that the overlap between training data and evaluation data was small enough to be ignored.

Results

Prescriptions

A total of 44,325,065 prescriptions from 22 million patients (138,535,092 records) at 349 hospitals, including 286 tertiary and 63 secondary hospitals, were used in our topic modeling process. This included 5,653 types of medications and 22,643 diseases or conditions. In the validation dataset, there were 14,166 (94.4%) outpatient prescriptions and 834 (5.6%) emergency prescriptions. Of these, 13,524 (90.2%) prescriptions

satisfied the appropriate criteria (marked as “appropriate”) and 1476 (9.8%) failed (marked as “inappropriate”).

Multiview Latent Dirichlet Allocation Topic Generation Results

By setting the topic parameters, we obtained eight topic models, all with commonly diagnosed diseases in clinical practice. For example, the model (K=30) included cardiovascular diseases, diabetes, chronic nephrosis, osteoporosis, and some respiratory infections. Regarding topic 27, hypertension had a 93.3% probability of appearing in this topic, and amlodipine, nifedipine, levamlodipine, and metoprolol had probabilities of 11.7%, 8.9%, 7.7%, and 6.1%, respectively. The top probability diagnoses in topic 23 were bronchitis, pneumonia, and bronchopneumonia, with the proportions of 55.6%, 21.8%, and 12.1%, respectively, whereas the corresponding medications were ambroxol (11.2%), budesonide (10.7%), azithromycin (9.9%), and terbutaline (6.2%). We also obtained topics related to gastrointestinal diseases and mental and dermal disorders. The details pertaining to the top 10 topics and their allocations are shown in Multimedia Appendix 2.

After comparing the training results of the topic models with settings at K=15, 20, 25, 30, 35, 40, 45, and 50 for the training results, it was found that a smaller topic number was associated with a weaker relation between the topics on one side and diagnoses and medications on the other, which were likely to appear more dispersed and had a lower probability of appearing in a topic. As the set value of the number of topics increased, the ability to summarize the disease was enhanced, that is, the subject-feature distribution of topic learning became more concentrated, the feature became more likely to appear in the topic, and the proportions of diagnosis and medication tended to be uniform.

Multiview Latent Dirichlet Allocation Evaluation Sensitivity Analysis Results

The BRPRD sample data evaluated the MV-LDA model. The performance of the MV-LDA model is shown in Figure 3. Each model showed higher specificity and NPV for some topics, with the NPV reaching more than 90%, and the sensitivity being the highest at a TM threshold of 1. As the threshold value declined, the sensitivity decreased, the specificity and PPV increased, and the NPV showed no relevant change. When the number of topics increased, the sensitivity increased greatly, but the specificity, PPV, and NPV changed little.

Taking all factors (sensitivity, specificity, PPV, NPV, and Youden’s index) into consideration, we set a cutoff of ≥1 TM anomaly scoring as the threshold for our MV-LDA detection model. The results showed a high sensitivity of 81.8% and a moderate specificity of 47.4%, and the PPV and NPV were 14.5% and 96.0%, respectively. These findings indicate that under the best performance parameter setting, we can find 1208 of 1476 inappropriate prescriptions.

Our model evaluation results revealed that the MV-LDA model had a better ability to detect inappropriate prescriptions when the TM threshold was set to 1. However, for a better understanding of the robustness of the results at this parameter setting, we performed a sensitivity analysis, repeating the experiments with separate sampling proportions of 90%, 70%, 50%, 30%, and 10%. Table 3 presents the findings. There were no relevant differences between the two experiments.

Figure 3. Summary of the performance of multiview latent Dirichlet allocation model with TM detection methods under different thresholds. Horizontal axis: thresholds of TM methods (from 1 to 5). Vertical axis: percentage of SEN, SPE, PPV, and NPV. K: number of topics; NPV: negative-predictive value; PPV: positive-predictive value; SEN: sensitivity; SPE: specificity; TM: topic mapping.

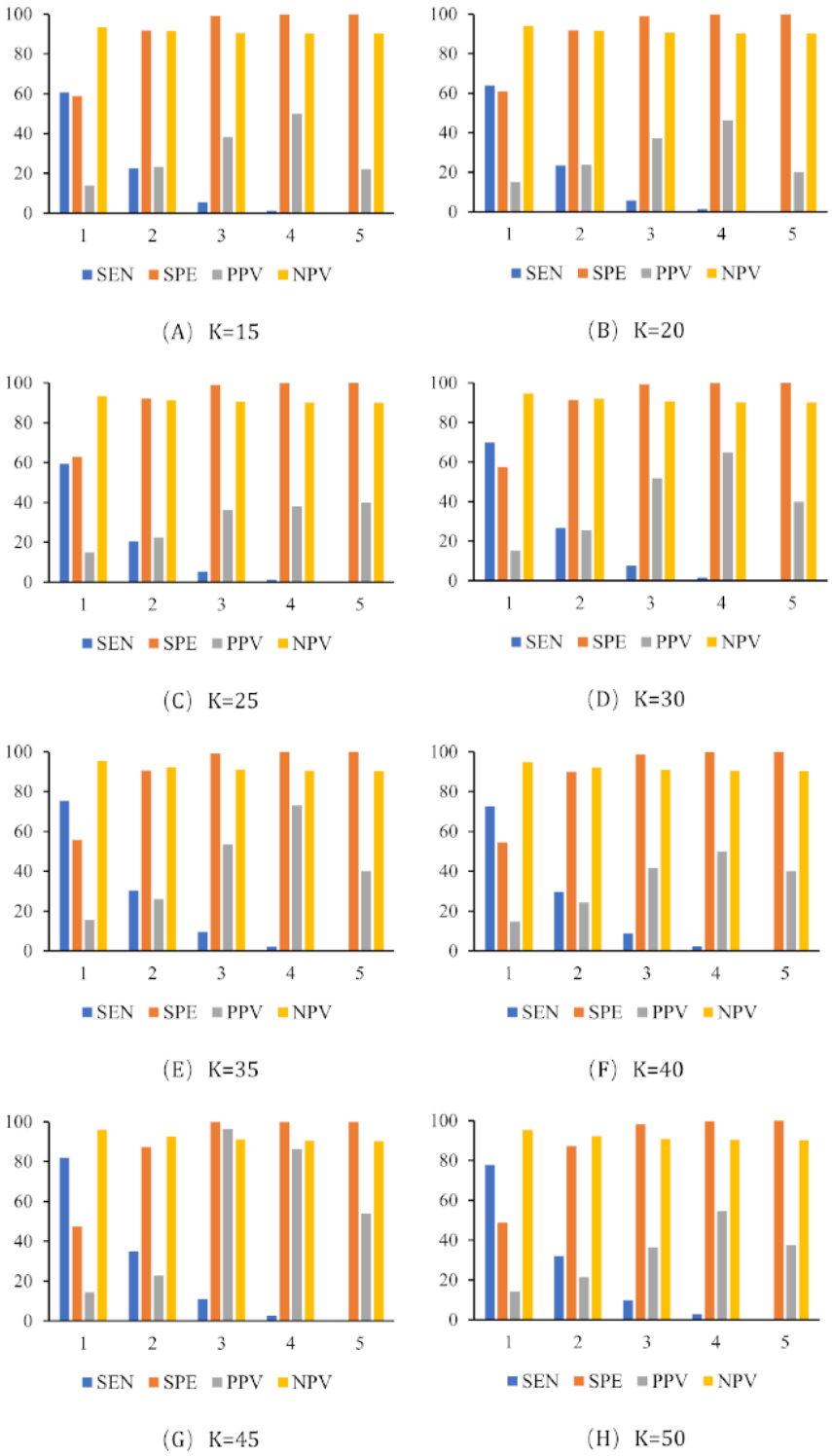


Table 3. Sensitivity analysis for multiview latent Dirichlet allocation with topic mapping detection methods (threshold=1).

Sampling proportion	TP[a], n (%)	FP[b], n (%)	FN[c], n (%)	TN[d], n (%)	SEN[e] (%)	SPE[f] (%)	PPV[g] (%)	NPV[h] (%)	Youden's index	AUROC[i]
90%	1073 (7.9)	5752 (42.6)	249 (1.8)	6435 (47.6)	81.2	47.2	14.3	95.9	14.3	0.689
70%	823 (7.9)	4446 (42.7)	179 (1.7)	4954 (47.6)	82.1	47.3	14.2	96.1	14.2	0.695
50%	570 (7.6)	3235 (43.1)	143 (1.9)	3562 (47.4)	79.9	47.6	13.8	95.8	13.8	0.686
30%	383 (8.6)	1958 (43.9)	93 (2.1)	2030 (45.5)	80.5	49.1	15.9	95.5	15.9	0.705
10%	109 (7.3)	622 (41.9)	21 (1.4)	734 (49.4)	83.8	45.9	12.9	96.7	12.9	0.693

[a]TP: true positive.

[b]FP: false positive.

[c]FN: false negative.

[d]TN: true negative.

[e]SEN: sensitivity.

[f]SPE: specificity.

[g]PPV: positive-predictive value.

[h]NPV: negative-predictive value.

[i]AUROC: area under the receiver operating characteristic curve.

Discussion

Principal Findings

The study drew upon the data of almost 45 million prescriptions obtained from the CMNRUD database (between October and December 2016). It then used the MV-LDA combination of TM anomaly detection and LDA topic modeling to build a model for detecting the inappropriate use of prescription medication. The model had a sensitivity of 81.8% and a specificity of 47.4% with 45 topics, and it had an anomaly threshold of 1 and showed stability in the sensitivity analysis. The topics that were already built into our study included most disorders, and the topics that were generated included noncommunicable diseases, such as cardiovascular diseases, which appeared in the largest proportions, consistent with clinical practice. The model also accommodated the tendency for many disorders to be seen in winter. For instance, upper respiratory infections, fever, and acute bronchitis were determined to be highly probable in winter. It took only 3.5 hours to generate the topics and only seconds to detect an anomaly, much quicker than the system of manual review or knowledge-based prescription review.

Limitations

The present work has several limitations. The MV-LDA model has more features than were used in this study, including usage, dosage, cost, and even laboratory test results, when available. Moreover, it was challenging to clean accessional variables from the data source. For example, there are multiple modes of recording dose packaging because the composition and dosage forms of medicines differ from each other. Because of the difficulties noted above, the first limitation is that our method ignored the medication's usage and dosage and only addressed the medication itself when building the MV-LDA model and validating the results. Second, the current model is still not supported for indicating the specific medication but tells us which prescriptions do not comply with the prescription review criteria. Besides, limited by the failure to obtain a labeled training dataset, the current model is not able to classify prescription misuse by criteria, such as the absence of proper indications, violation of clinical guidelines, and misuse of dosage, and can only detect the appropriateness of prescriptions. This study is also limited by the diverse structure of the model training and evaluation database and a minor overlap between the datasets used. However, we thought that a minor overlap of the data might not be associated with a major change in the results. Meanwhile, the two databases showed a commonality in their treatment patterns, and this, in fact, could be a topic for exploratory research on methodology. While this study focused on the development of a model, in the future, we will address a diverse range of parameters to determine the most effective MV-LDA model for detecting prescription misuse.

Comparison With Prior Work

Studies, such as those encouraging regular medication review and introducing automated information systems [47,48], have been conducted with the aim of controlling the inappropriate use of medications in China. However, the increasing number of new drugs entering the market, delays in updating the databases, and insufficient knowledge of medications all raise the probability of nonideal use [49]. Knowledge-based and experience-based software has relevant limitations, including efficiency constraints. However, data mining techniques are customizable and can identify inappropriate prescriptions. For example, association rule mining has been used to find inappropriate prescriptions by calculating the co-occurrence of medications and diseases, resulting in a sensitivity of 75.9% and a specificity of 89.5% [42,43]. These methods however do have disadvantages, including inefficiency in the generation of candidate item sets because they require vast data sources and the frequent scanning of databases.

Furthermore, these methods often fail to explore latent structures and are prone to making spurious associations that can mislead clinical practitioners. Besides, in a previous study, a model combining natural language processing with guidelines based on expert knowledge was used to detect medication overuse,

and it showed degrees of sensitivity and specificity that were similar to those in our study [50]. Despite using different data sources and operating under diverse study conditions, we noted a higher sensitivity as compared with the association rule mining method. Although we failed to obtain a higher PPV, which is strongly related to the prevalence of inappropriate prescriptions, we think the MV-LDA model is suitable for preliminary screening and can be an alternative detection method, allowing clinical practice to flag potentially inappropriate prescriptions for manual review. Such a step could save a large amount of working time and reduce labor intensity.

A topic model is a multiple machine learning method and is used to reveal the semantics in the body of the text. With its advantages of topic extraction and model expansibility, LDA has become a commonly used topic modeling method. It was first used to extract underlying semantics and was then optimized to become a robust means of text mining analysis for social media [51-53]. Recently, topic modeling methods, particularly LDA, have been used for both structured and unstructured clinical data [21,26,54-56]. Several studies have attempted to scale the efficiency of LDA's topic generation. The symptom-herb-diagnosis topic model, which was proposed to determine the association between treatment with Chinese medicine and diabetes, can be used to find herbs to treat specific symptoms [55]. Multiple-channel LDA [57] focused on the support system for clinical decisions and based itself on a similar concept that the coupling of diagnoses and medications reflects the health status of patients at the time of seeing a doctor. However, we were not able to obtain a piece of well-recorded contextual information or additional information, but we could leverage two variables (diagnosis and medication) and realize the aim of the study. Besides, given the shortcomings of miscellaneous algorithms and more extended calculations of an LDA-based model, various measures were taken to improve the MV-LDA algorithm in our project, which is not the focus of this study. The multiview topic modeling approach used here has been previously tested in different languages [34]. This allowed us to take both medications and diagnoses into consideration simultaneously. We leveraged these advantages and processed only half of the data used in a previous study to determine the association between diagnoses and medications for each prescription [58]. Prescriptions in our study were considered the equivalent of articles in that previous study, reflecting the particular situation of those patients. The topics and their allocations were consistent with clinical practice, providing proof of the robustness of our method.

Conclusions

Our MV-LDA model can train the distribution of diagnosis-medication topics from a large number of prescriptions and can detect the potentially inappropriate use of prescription medications when combined with the TM method. Considering its mediocre specificity and moderate sensitivity, this model can be used as a primary screening tool and will likely complement and improve manual review. The model still needs more extension of views (introduction of more variables) to make full use of the information in the prescription and further improve the ability to identify prescription misuse.

Acknowledgments

This work was supported by the National Natural Foundation of China (grant number 91646107) and the Beijing Municipal Science and Technology Project (grant number D151100002215002).

Authors' Contributions

LZ was the principal investigator for this study; she contributed to the design, analysis, and interpretation of the study. YY and YC contributed to the analysis and interpretation of data and provided clinical support. SL, ST, and JZ contributed to the model's development. YH and JZ contributed to data extraction. LZ and YC drafted the manuscript. SZ provided overall supervision of the study and critically edited the manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of this version for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Confusion matrix of model evaluation (multiview latent Dirichlet allocation model versus experts).
[DOCX File , 14 KB - medinform_v8i7e16312_app1.docx]

Multimedia Appendix 2
Example results of multiview latent Dirichlet allocation topic modeling (K=30).
[DOCX File , 25 KB - medinform_v8i7e16312_app2.docx]

References

1. World Health Organization. 2004. World Medicines Situation URL: https://www.who.int/medicines/areas/policy/world_medicines_situation/en/ [accessed 2019-11-14]

2. Holloway K, Dijk LV. World Health Organization. 2011. The World Medicines Situation 2011: Rational Use of Medicines URL: https://www.who.int/medicines/areas/policy/world_medicines_situation/WMS_ch14_wRational.pdf [accessed 2019-10-12]
3. Hamilton HJ, Gallagher PF, O'Mahony D. Inappropriate prescribing and adverse drug events in older people. BMC Geriatr 2009 Jan 28;9:5 [FREE Full text] [doi: 10.1186/1471-2318-9-5] [Medline: 19175914]
4. Lederberg J. Infectious history. Science 2000 Apr 14;288(5464):287-293. [doi: 10.1126/science.288.5464.287] [Medline: 10777411]
5. Reynolds L, McKee M. Factors influencing antibiotic prescribing in China: an exploratory analysis. Health Policy 2009 Apr;90(1):32-36. [doi: 10.1016/j.healthpol.2008.09.002] [Medline: 18849089]
6. Song Y, Bian Y, Petzold M, Li L, Yin A. The impact of China's national essential medicine system on improving rational drug use in primary health care facilities: an empirical study in four provinces. BMC Health Serv Res 2014 Oct 25;14:507 [FREE Full text] [doi: 10.1186/s12913-014-0507-3] [Medline: 25344413]
7. Li Y, Xu J, Wang F, Wang B, Liu L, Hou W, et al. Overprescribing in China, driven by financial incentives, results in very high use of antibiotics, injections, and corticosteroids. Health Aff (Millwood) 2012 May;31(5):1075-1082. [doi: 10.1377/hlthaff.2010.0965] [Medline: 22566449]
8. Ministry of Health of the People's Republic of China. 2010. Management Practices of Hospital Prescription Comment (trial) URL: http://www.gov.cn/gzdt/2010-03/04/content_1547080.htm [accessed 2018-02-06]
9. Zhang Y, Li P, Li J, Wang D, Mei D, Zhang B. Influence of automated pharmacy system on waiting time in outpatient pharmacy. Chinese Journal of Hospital Pharmacy (1) 2014:63-66. [doi: 10.13286/j.cnki.chinhosppharmacyj.2014.01.19]
10. Peking University Third Hospital. 2018. Outpatient pharmacy URL: https://www.puh3.net.cn/yjk/ksbm/153254.shtml [accessed 2018-11-04]
11. Wang X, Gong Z, Zhou Y, Huang Y. Research of real-time review system of electronic prescriptions for putpatient and emergency in hospital. Science Mosaic 2016(5):33-35.
12. Gao Y, Fu L, Zhong X, Liu Z. Discussions on Problems about the Monitoring System for Rational Drug Use and Relevant Countermeasures. China Pharmacy 2015(22):3159-3161. [doi: 10.6039/j.issn.1001-0408.2015.22.42]
13. Meyer J, Ostrzinski S, Fredrich D, Havemann C, Krafczyk J, Hoffmann W. Efficient data management in a large-scale epidemiology research project. Comput Methods Programs Biomed 2012 Sep;107(3):425-435. [doi: 10.1016/j.cmpb.2010.12.016] [Medline: 21256617]
14. Peikari M, Salama S, Nofech-Mozes S, Martel AL. A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. Sci Rep 2018 May 08;8(1):7193 [FREE Full text] [doi: 10.1038/s41598-018-24876-0] [Medline: 29739993]
15. Hu X, Gallagher M, Loveday W, Connor J, Wiles J. Detecting anomalies in controlled drug prescription data using probabilistic models. : Springer, Cham; 2015 Presented at: Australasian Conference on Artificial Life and Computational Intelligence; February 5-7, 2015; Newcastle, NSW, Australia p. 337-349. [doi: 10.1007/978-3-319-14803-8_26]
16. Nirad D, Surendro K. Outlier detection using association rule mining for information quality improvement. 2017 Presented at: International Conference on Recent Trends in Science, Engineering and Technology; July 10-11, 2017; Bangkok, Thailand. [doi: 10.17758/eap.dir0717002]
17. Valko M, Kveton B, Valizadegan H, Cooper G, Hauskrecht M. Conditional anomaly detection with soft harmonic functions. : IEEE; 2011 Presented at: 11th International Conference on Data Mining; December 11-14, 2011; Vancouver, BC, Canada p. 735-743. [doi: 10.1109/icdm.2011.40]
18. Song X, Wu M, Jermaine C, Ranka S. Conditional Anomaly Detection. IEEE Trans Knowl Data Eng 2007 May;19(5):631-645. [doi: 10.1109/tkde.2007.1009]
19. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 02;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
20. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3(4-5):993-1022. [doi: 10.1056/NEJM196803212781204] [Medline: 5637250]
21. Park S, Choi D, Kim M, Cha W, Kim C, Moon I. Identifying prescription patterns with a topic model of diseases and medications. J Biomed Inform 2017;75:35-47 [FREE Full text] [doi: 10.1016/j.jbi.2017.09.003] [Medline: 28958484]
22. Zeng QT, Redd D, Rindflesch T, Nebeker J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. AMIA Annu Symp Proc 2012;2012:1050-1059 [FREE Full text] [Medline: 23304381]
23. Shivashankar S, Srivathsan S, Ravindran B, Tendulkar AV. Multi-view methods for protein structure comparison using latent dirichlet allocation. Bioinformatics 2011 Jul 01;27(13):i61-i68 [FREE Full text] [doi: 10.1093/bioinformatics/btr249] [Medline: 21685102]
24. Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H. Discovery of clinical pathway patterns from event logs using probabilistic topic models. J Biomed Inform 2014 Feb;47:39-57 [FREE Full text] [doi: 10.1016/j.jbi.2013.09.003] [Medline: 24076435]
25. Huang Z, Dong W, Ji L, He C, Duan H. Incorporating comorbidities into latent treatment pattern mining for clinical pathways. J Biomed Inform 2016 Feb;59:227-239 [FREE Full text] [doi: 10.1016/j.jbi.2015.12.012] [Medline: 26719169]
26. Zhang L, Zhao J, Wang Y, Xie B. Mining Patterns of Disease Progression: A Topic-Model-Based Approach. Stud Health Technol Inform 2016;228:354-358. [Medline: 27577403]

27. Chong W, Blei D, Li FF. Simultaneous image classification and annotation. 2009 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL, USA p. 1903-1910. [doi: 10.1109/cvpr.2009.5206800]
28. Wang X, Ma X, Grimson E. Unsupervised activity perception by hierarchical bayesian models. IEEE Trans Pattern Anal Mach Intell 2009;31(3):539-555. [doi: 10.1109/cvpr.2007.383072]
29. Poldrack RA, Mumford JA, Schonberg T, Kalar D, Barman B, Yarkoni T. Discovering relations between mind, brain, and mental disorders using topic mapping. PLoS Comput Biol 2012;8(10):e1002707 [FREE Full text] [doi: 10.1371/journal.pcbi.1002707] [Medline: 23071428]
30. Huang Z, Dong W, Duan H. A probabilistic topic model for clinical risk stratification from electronic health records. J Biomed Inform 2015 Dec;58:28-36 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.005] [Medline: 26370451]
31. Liu B, Liu L, Tsykin A, Goodall G, Green J, Zhu M, et al. Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. Bioinformatics 2010 Dec 15;26(24):3105-3111 [FREE Full text] [doi: 10.1093/bioinformatics/btq576] [Medline: 20956247]
32. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. J Biomed Inform 2015 Dec;58:156-165 [FREE Full text] [doi: 10.1016/j.jbi.2015.10.001] [Medline: 26464024]
33. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus 2016;5(1):1608 [FREE Full text] [doi: 10.1186/s40064-016-3252-8] [Medline: 27652181]
34. Zhang G, Iwata T, Kashima H. Robust multi-view topic modeling by incorporating detecting anomalies. 2017 Presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; September 18-22, 2017; Skopje, Macedonia p. 238-250. [doi: 10.1007/978-3-319-71246-8_15]
35. Sun S. A survey of multi-view machine learning. Neural Comput & Applic 2013 Feb 17;23(7-8):2031-2038. [doi: 10.1007/s00521-013-1362-6]
36. Xu C, Tao D, Xu C. A survey on multi-view learning. arXiv preprint 2013 [FREE Full text]
37. Zhang L, Li X, Liu H, Mei J, Hu G, Zhao J. Probabilistic-mismatch anomaly detection: do one's medications match with the diagnoses. 2016 Presented at: IEEE 16th International Conference on Data Mining (ICDM); December 12-15, 2016; Barcelona, Spain p. 659-668. [doi: 10.1109/icdm.2016.0077]
38. Hu Y. Establishment and application of Chineses monitoring network for the rational use of drugs system. Exploration of Rational Drug Use in China 2009;6(8):5-8.
39. Yang Y, Zhou X, Gao S, Lin H, Xie Y, Feng Y, et al. Evaluation of Electronic Healthcare Databases for Post-Marketing Drug Safety Surveillance and Pharmacoepidemiology in China. Drug Saf 2018 Jan;41(1):125-137. [doi: 10.1007/s40264-017-0589-z] [Medline: 28815480]
40. Zhen J, Bian B, Kong F, Yan B. Effect evaluation of regional prescription review on rational clinical drug use. Chinese Journal of Hospital Administration 2015(7):531-533.
41. Ministry of Health of the People's Republic of China. Prescription Administrative Policy 2007 URL: http://www.gov.cn/flfg/2007-03/13/content_549406.htm [accessed 2019-11-08]
42. Nguyen PA, Syed-Abdul S, Iqbal U, Hsu M, Huang C, Li H, et al. A probabilistic model for reducing medication errors. PLoS One 2013;8(12):e82401 [FREE Full text] [doi: 10.1371/journal.pone.0082401] [Medline: 24312659]
43. Yang H, Iqbal U, Nguyen PA, Lin S, Huang C, Jian W, et al. An automated technique to identify potential inappropriate traditional Chinese medicine (TCM) prescriptions. Pharmacoepidemiol Drug Saf 2016 Apr;25(4):422-430. [doi: 10.1002/pds.3976] [Medline: 26910512]
44. Yang M, Wang D, Wang X, Zhang Y. Prescription review and inappropriate prescription analysis in our hospital in 2013. Chinese Medical Science 2014(16):129-131.
45. Li DC, Thermeau T, Chute C, Liu H. Discovering associations among diagnosis groups using topic modeling. AMIA Jt Summits Transl Sci Proc 2014;2014:43-49 [FREE Full text] [Medline: 25954576]
46. Liu S, Tang S, Zhao J, Wang Y, Zhuo L. An extended topic model based abnormal medical prescription detection method. 2018 Presented at: National Conference on Pervasive Computing; September 14-16, 2018; Tianjin, China.
47. World Health Organization. 2016. Medication Errors: Technical Series on Safer Primary Care URL: https://apps.who.int/iris/bitstream/handle/10665/252274/9789241511643-eng.pdf?sequence=1 [accessed 2020-04-24]
48. Velo GP, Minuz P. Medication errors: prescribing faults and prescription errors. Br J Clin Pharmacol 2009 Jun;67(6):624-628 [FREE Full text] [doi: 10.1111/j.1365-2125.2009.03425.x] [Medline: 19594530]
49. Yuan N, Chen N. Reasons of irrational drug use in medical institutions and its countermeasures: a case study of irrational drug use in department of gastroenterology. Medicine and Philosophy 2015;36(15):51-53.
50. Salmasian H, Freedberg DE, Abrams JA, Friedman C. An automated tool for detecting medication overuse based on the electronic health records. Pharmacoepidemiol Drug Saf 2013 Feb;22(2):183-189 [FREE Full text] [doi: 10.1002/pds.3387] [Medline: 23233423]
51. Zhao W, Jiang J, Weng J, He J, Lim E, Yan H, et al. Comparing twitter and traditional media using topic models. 2011 Presented at: European Conference on Information Retrieval; April 18-21, 2011; Dublin, Ireland p. 338-349. [doi: 10.1007/978-3-642-20161-5_34]

52. Xianghua F, Guo L, Yanyan G, Zhiqiang W. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowledge-Based Systems 2013 Jan;37(2):186-195. [doi: 10.1016/j.knosys.2012.08.003]
53. Fu X, Liu G, Guo Y, Guo W. Multi-aspect blog sentiment analysis based on LDA topic model and Hownet Lexicon. 2011 Presented at: International Conference on Web Information Systems and Mining; September 24-25, 2011; Taiyuan, China p. 131-138. [doi: 10.1007/978-3-642-23982-3_17]
54. Lin F, Xiahou J, Xu Z. TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model. Multimed Tools Appl 2016 Apr 13;75(22):14203-14232. [doi: 10.1007/s11042-016-3363-9]
55. Zhang X, Zhou X, Huang H, Feng Q, Chen S, Liu B. Topic model for Chinese medicine diagnosis and prescription regularities analysis: case on diabetes. Chin J Integr Med 2011 Apr;17(4):307-313. [doi: 10.1007/s11655-011-0699-x] [Medline: 21509676]
56. Cohen R, Aviram I, Elhadad M, Elhadad N. Redundancy-aware topic modeling for patient record notes. PLoS One 2014;9(2):e87555 [FREE Full text] [doi: 10.1371/journal.pone.0087555] [Medline: 24551060]
57. Lu H, Wei C, Hsiao F. Modeling healthcare data using multiple-channel latent Dirichlet allocation. J Biomed Inform 2016 Apr;60:210-223 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.003] [Medline: 26898516]
58. Nguyen C, Zhan D, Zhou Z. Multi-modal image annotation with multi-instance multi-label LDA. 2013 Presented at: 23rd International Joint Conference on Artificial Intelligence; August 3-9, 2013; Beijing, China.

Abbreviations

BRPRD: Beijing Regional Prescription Review Database
CMNRUD: Chinese Monitoring Network for the Rational Use of Drugs
LDA: latent Dirichlet allocation
MV-LDA: multiview latent Dirichlet allocation
NPV: negative-predictive value
PPV: positive-predictive value
TM: topic mapping

Edited by G Eysenbach; submitted 18.09.19; peer-reviewed by Z Yang, A Aminbeidokhti; comments to author 25.11.19; revised version received 18.01.20; accepted 24.03.20; published 06.07.20.

Please cite as:
Zhuo L, Cheng Y, Liu S, Yang Y, Tang S, Zhen J, Zhao J, Zhan S
A Multiview Model for Detecting the Inappropriate Use of Prescription Medication: Machine Learning Approach
JMIR Med Inform 2020;8(7):e16312
URL: https://medinform.jmir.org/2020/7/e16312
doi:10.2196/16312
PMID:32209527

Original Paper

Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation

Zhenzhen Du[1,2*], MSc; Yujie Yang[1,3*], MSc; Jing Zheng[4*], PhD; Qi Li[1], MSc; Denan Lin[4], MSc; Ye Li[1], PhD; Jianping Fan[1], PhD; Wen Cheng[2], PhD; Xie-Hui Chen[5], MSc; Yunpeng Cai[1], PhD

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[2]Fiberhome Technologies College, Wuhan Research Institute of Posts and Telecommunications, Wuhan, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]Shenzhen Health Information Center, Shenzhen, China

[5]FuWai Hospital, Chinese Academy of Medical Sciences, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Yunpeng Cai, PhD
Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences
1068 Xueyuan Blvd
Nanshan District
Shenzhen
China
Phone: 86 755 86392202
Fax: 86 755 86392299
Email: yp.cai@siat.ac.cn

Abstract

Background: Predictions of cardiovascular disease risks based on health records have long attracted broad research interests. Despite extensive efforts, the prediction accuracy has remained unsatisfactory. This raises the question as to whether the data insufficiency, statistical and machine-learning methods, or intrinsic noise have hindered the performance of previous approaches, and how these issues can be alleviated.

Objective: Based on a large population of patients with hypertension in Shenzhen, China, we aimed to establish a high-precision coronary heart disease (CHD) prediction model through big data and machine-learning

Methods: Data from a large cohort of 42,676 patients with hypertension, including 20,156 patients with CHD onset, were investigated from electronic health records (EHRs) 1-3 years prior to CHD onset (for CHD-positive cases) or during a disease-free follow-up period of more than 3 years (for CHD-negative cases). The population was divided evenly into independent training and test datasets. Various machine-learning methods were adopted on the training set to achieve high-accuracy prediction models and the results were compared with traditional statistical methods and well-known risk scales. Comparison analyses were performed to investigate the effects of training sample size, factor sets, and modeling approaches on the prediction performance.

Results: An ensemble method, XGBoost, achieved high accuracy in predicting 3-year CHD onset for the independent test dataset with an area under the receiver operating characteristic curve (AUC) value of 0.943. Comparison analysis showed that nonlinear models (K-nearest neighbor AUC 0.908, random forest AUC 0.938) outperform linear models (logistic regression AUC 0.865) on the same datasets, and machine-learning methods significantly surpassed traditional risk scales or fixed models (eg, Framingham cardiovascular disease risk models). Further analyses revealed that using time-dependent features obtained from multiple records, including both statistical variables and changing-trend variables, helped to improve the performance compared to using only static features. Subpopulation analysis showed that the impact of feature design had a more significant effect on model accuracy than the population size. Marginal effect analysis showed that both traditional and EHR factors exhibited highly nonlinear characteristics with respect to the risk scores.

Conclusions: We demonstrated that accurate risk prediction of CHD from EHRs is possible given a sufficiently large population of training data. Sophisticated machine-learning methods played an important role in tackling the heterogeneity and nonlinear nature of disease prediction. Moreover, accumulated EHR data over multiple time points provided additional features that were valuable for risk prediction. Our study highlights the importance of accumulating big data from EHRs for accurate disease predictions.

(JMIR Med Inform 2020;8(7):e17257) doi:10.2196/17257

KEYWORDS

coronary heart disease; machine learning; electronic health records; predictive algorithms; hypertension

Introduction

Cardiovascular diseases (CVDs) are currently the primary cause of global deaths according to a survey from the World Health Organization [1]. In 2016, 17.9 million people were estimated to have died of CVDs, representing 31% of all global deaths. Among these deaths, 85% are due to heart attack and stroke [2]. Modeling and prediction of CVD risk have long attracted the interest of many researchers. Several well-known risk scales such as the Framingham scales [3-5], American College of Cardiology/American Heart Association scales [6], QRISK [7], QRISK2 [8], and SCORE [9] have been established following years of population cohort studies, which provide an effective reference for clinicians to carry out disease prevention and treatment work [10].

Nevertheless, due to the complex and heterogeneous nature of CVD pathology, the prediction power of these risk scales has proven to be rather limited [11,12]. In recent years, researchers have been discovering or proposing new risk factors of CVDs according to lifestyle [13-15]; biochemical testing [16-18]; electrocardiograms [19-22]; medical imaging [23-28]; genetic, genomic, and proteomic biomarkers [29,30]; along with microbe and gene-environment interactions [31]. The steady growth of new emerging risk biomarkers surges demands for developing more precise disease prediction models. However, the traditional paradigm used for building risk models from a population-based study imposes a severe challenge to the development of accurate risk models, which usually requires a fixed set of observation variables at the beginning of the study and a lengthy follow-up period to collect all outcomes. Moreover, recent studies have identified that CVD risk factors vary according to social environments as well as ethnic and geographic differences [32,33]. This implies that an adaptive approach should be adopted for constructing more accurate CVD risk models that can be tuned to a specific population with higher efficiency.

Recently, the boosting of national or region-wide electronic health record (EHR) management systems has enabled the sharing and fusion of EHR data from many institutes [34], providing a faster approach for collecting large-scale population data to carry out retrospective cohort studies for more efficient assessments of CVD risk factors. A large-scale follow-up study using the EHR data of 1.25 million people identified the heterogeneous associations of blood pressure across different CVDs and age groups [35], which could not be discovered in previous population studies. Several efforts have also been made to create new disease risk prediction models based on EHR data using machine-learning models such as logistic regression, support vector machine (SVM), or K-nearest neighbor (KNN) approaches [36-39], but most of the results demonstrated very limited advantages compared with traditional risk scales. Compared with traditional cohort studies, EHR data are easier to acquire but the data quality is significantly inferior. Hence, one question that arises is whether EHR data are intrinsically unreliable and therefore unsuitable for achieving high-accuracy predictions. Moreover, studies on machine-learning approaches in EHR-based risk modeling are rather limited in the sense that almost all of the methods reported to date involve converting the EHR data into a single matrix, resulting in a lack of dynamic information. Therefore, establishment of a better modeling technique, more advanced machine-learning methods, and more data resources are expected to provide positive contribution to the power of existing prediction models.

Toward this end, the aim of the present study was to address these issues based on a case study using a large population of registered patients with hypertension in Shenzhen, China. Specifically, we evaluated the possibility of establishing a high-precision coronary heart disease (CHD) prediction model through big data and machine-learning methods. With a large population of 20,156 patients with CHD onset and more than 100 original features gathered from EHRs accumulated over 8 years, we were able to obtain more insight into risk factors than possible with traditional cohort studies, demonstrating that accurate prediction of CHD risks could be possible with the aid of large datasets, sophisticated machine-learning methods, and dynamic trends of patient information extracted from multiple time-point EHR records. These findings highlight the importance of accumulating EHR big data for accurate disease risk modeling, and provide a useful approach for the early screening and prevention of CVDs.

Methods

Overview of Sample and Data Processing

We investigated the stocked EHRs of registered patients with hypertension from the Shenzhen Health Information platform, which gathered the clinical records of 83 local public hospitals and over 600 community health service centers from 2010 to 2018. Each patient visiting the associated hospitals was assigned a unique identifier so that the clinical activities at multiple institutes could be merged. De-identification was performed on all data by the platform administrators under supervision of the Shenzhen Municipal Health Commission before collecting the datasets for investigation. Since all of the data were collected during regular clinical activities and were anonymized, following the Guidelines of the World Medical Association's

Declaration of Helsinki term 32, a waive-of-consent protocol was adopted, which was approved by the Shenzhen Institutes of Advanced Technology Institutional Review Board (No. SIAT-IRB-151115-H0084).

A total of 251,791 registered patients with hypertension were identified in the platform data. The collected EHR data for each patient included regular chronic disease follow-up records, inpatient and outpatient records, and clinical examinations and biochemical tests. Detailed field descriptions are provided in Multimedia Appendix 1. CHD diagnosis results were extracted from the main diagnosis field of the inpatient or outpatient records using the International Statistical Classification of Diseases and Related Health Problems (ICD)-10 [40] diagnostic codes I20 to I25 or the keywords related to CHD conditions, including “coronary heart disease,” “coronary sclerosis heart disease,” “ischemic cardiomyopathy,” “angina,” “acute myocardial infarction,” “myocardial ischemia,” “heart failure” (all translated from Chinese), and others, resulting in 37,776 cases of CHD onset.

To ensure the reliability of the outcomes, we required all samples to be associated with regular chronic disease follow-up information. A total of 23,335 samples were thus removed, resulting in 228,456 samples for analysis. We defined the follow-up period for each patient as the time interval between the most recent and the earliest record (regardless of record types) collected in the system. For positive samples (patients with CHD onset, n=33,279), we required the patient to be CHD free at the initial state and for the interval between the time of CHD diagnosis and the last CHD-free follow-up time to be within 0-3 years, which excluded 9027 patients, leaving 24,252 patients. Among the excluded patients, 9018 had a diagnosis of CHD onset but the diagnosis time was more than 3 years after the latest CHD-free follow-up. To avoid possible latencies in diagnosis, we excluded these patients from the present analysis, but the distribution of their prediction scores was analyzed later. For negative samples (non-CHD patients, n=195,177), we excluded 23,054 patients with other severe diseases (eg, death, stroke, cancer/tumor, renal failure, rheumatic heart disease, pulmonary heart disease, pericardial defect, heart valve disease, congestive heart failure, acute myocardial infarction) and 120,717 patients with a follow-up period less than 3 years, resulting in a set of 51,606 non-CHD samples. The reason for excluding patients with heart failure and myocardial infarction from the non-CHD set was that there may be a suspicion of CHD in such cases but without an explicit diagnosis. In addition, patients with other severe diseases would receive intensive medical interventions; thus, some of these patients may have previously had cardiac risks but interventions were administered prior to making a diagnosis of CHD. For example, the CHD risk scores of stroke patients without CHD were predicted to be high using our model (Multimedia Appendix 2); hence, these cases were excluded to avoid confusion. For positive samples, only the records during the CHD-free period were used for investigation. For negative samples, only the records from at least 3 years before the study endpoint were included. The recording time of the most recent included record for each patient was assigned as the baseline time point.

EHR data usually contain abundant missing values. To avoid the influence of missing data on the prediction results, we used four basic variables as the quality filter of samples: age, gender, systolic blood pressure, and hypertension diagnosis time. Samples with no valid values for any of the above variables were excluded from the analysis. Moreover, only patients aged between 20 to 85 years were included in the study. Finally, we included data for 42,676 patients in the research cohort who met the above conditions, comprising 20,156 patients with CHD and 22,520 non-CHD patients. The above pipeline is schematically presented in Figure 1. Finally, the positive and negative samples were divided evenly to form the training set and the test set, respectively. Table 1 and Table 2 summarize the basic characteristics of both datasets. The distribution of the CHD-free time for the CHD group is shown in Multimedia Appendix 3.

Figure 1. Patient cohort data processing. CHD: coronary heart disease; EHR: electronic health record.

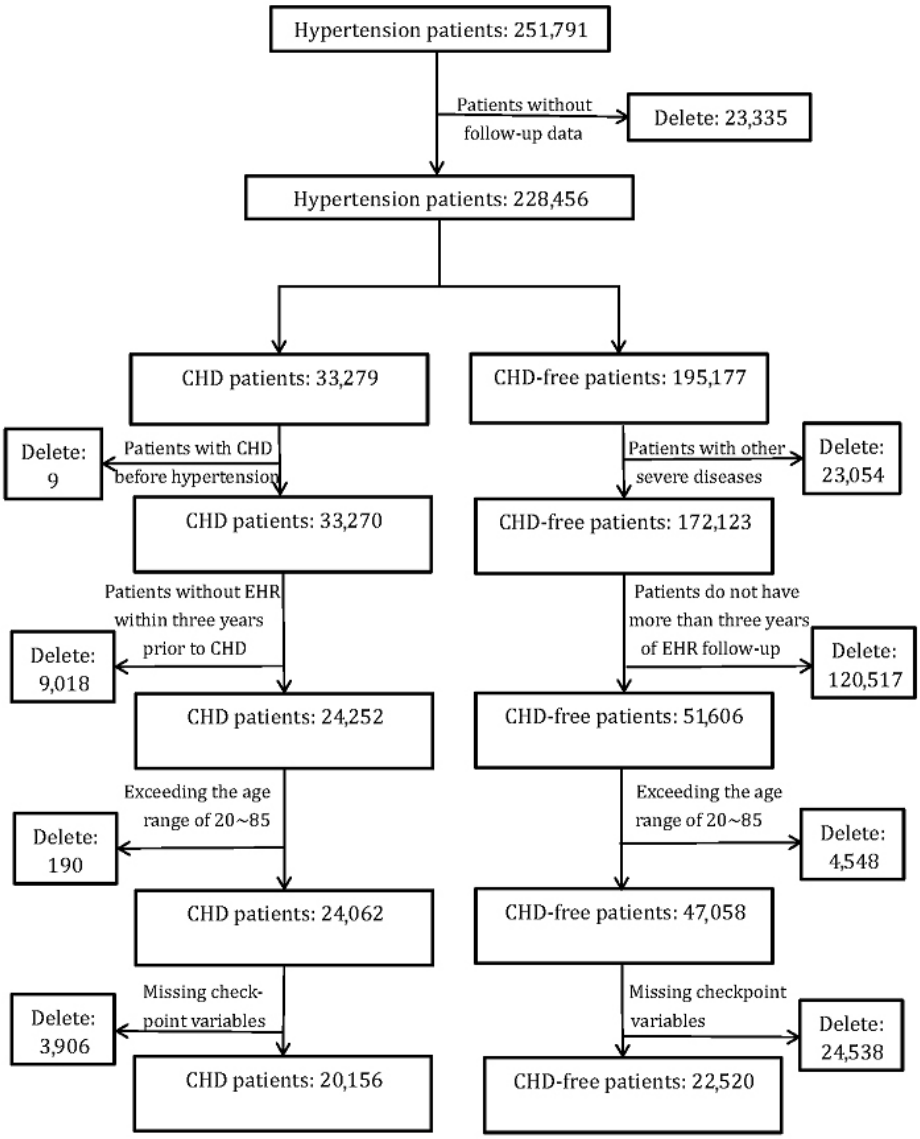


Table 1. Sample distribution of the training and test datasets.

Subsample	Training set (N=21,338), n (%)	Test set (N=21,338), n (%)
Males	12,303 (57.66)	12,286 (57.58)
Females	9035 (42.34)	9052 (42.42)
Positive samples	10,078 (47.23)	10,078 (47.23)
Negative samples	11,260 (52.77)	11,260 (52.77)

Table 2. Basic characteristics of subjects for the two datasets.

Characteristic	Training set (N=21,338), mean (SD)	Test set (N=21,338), mean (SD)
Duration of illness (years)	5.8 (5.10)	5.7 (5.01)
Age (years)	49.97 (12.01)	49.52 (11.99)
Last SBP[a] (mmHg)	131.39 (10.49)	131.37 (10.37)
Maximum SBP (mmHg)	135.40 (11.85)	135.70 (12.01)
Minimum SBP (mmHg)	128.21 (10.45)	127.96 (10.49)
Mean SBP (mmHg)	131.68 (9.66)	131.69 (9.57)

[a]SBP: systolic blood pressure.

Feature Processing

In contrast to most existing research in the field, our dataset included multiple records with different record times for each patient. Therefore, data preprocessing and feature variable extraction, selection, and construction were crucial steps for the establishment and analysis of our model.

First, variables with over 20% missing values were removed from the study. Second, text parsing was performed. Inpatient and outpatient diagnostic results are a mixture of ICD codes and natural language text input. If the ICD codes were available for a record, we used the ICD codes directly as the annotation or features of the samples. Otherwise, by using an inhouse-designed lexical parsing code with keyword mapping and error corrections, we converted the diagnostic text into corresponding ICD codes. The parser was rule-based, in which each ICD code item was mapped to varied texts through a regular expression of keywords. The parsing procedure was carried out iteratively. At the end of each loop, the unparsed texts were collected and sorted by word frequency, and then a manual inspection was performed and the expressions were modified to match more text (including tolerating typographical errors). The loops continued until the unparsed texts were considered noninformative.

Third, accounting was carried out. Features from multiple sources (eg, examination, inpatient, and outpatient records) or multiple time points representing the same physiology index were gathered, and their maximum, minimum, or average values were calculated and used as new features. Fourth, for some rare diagnostic symptoms and similar symptoms (eg, diseases belonging to the same ICD class but less related to cardiac events) were merged into a single variable to avoid sparsity in value distribution. Finally, we divided the follow-up period of each patient into the early and late halves at the mid-time points. The frequency of specified events (eg, in-hospital or out-hospital visits, symptom onset) were accounted for each half, and the ratios were used as a new variable representing the trending status of the patients.

Machine-Learning Algorithms

Extreme Gradient Boosting

Our model is based on the machine-learning algorithm XGBoost [41], which is short for extreme gradient boosting approach. XGBoost is an integrated machine-learning algorithm based on multiple decision trees with gradient boost as the framework. The loss function of XGBoost is defined as follows:

Where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}i$ and the target y_i. The second term Ω, as a regularization term, penalizes the complexity of the model. In contrast to the traditional gradient boosting decision tree method, XGBoost performs a second-order Taylor expansion on the loss function, and the additional regularization term helps to find the optimal solution for the whole, followed by weighing the decline of the objective function and the complexity of the model to avoid overfitting [41].

XGBoost supports missing values by default and naturally accepts a sparse feature format, allowing for directly feeding the data as a sparse matrix, and only contains nonmissing values (ie, features that are not presented in the sparse feature matrix are treated as “missing” and XGBoost will handle them internally). In tree algorithms, branch directions for missing values are learned during training. Internally, XGBoost treats nonpresence as a missing value and learns the best direction to handle missing values [41]. Equivalently, this can be viewed as automatically “learning” the best imputation values based on loss reduction. For continuous features, a missing (default) direction is learnt for missing value data to go into, so that missing data of a specific value will go in the default direction.

SVM

SVM is a generalized linear classifier that classifies data in a supervised learning manner, which was developed by Cortes and Vapnik [42]. The decision boundary is the maximum-margin hyperplane that solves the learning sample. The model trains a function that calculates a score for a new input to separate samples into two classes by building this hyperplane [43].

Logistic Regression

Logistic regression is a generalized linear regression analysis model [44], which is often used in data mining, automatic disease diagnosis, economic forecasting, and other broad applications. The algorithm is essentially a common two-category model, and the category corresponding to the object is obtained by inputting the attribute sequence of the object. The model assumes that the data obey the Bernoulli distribution, and uses the method of maximizing the likelihood

function to solve the parameters with gradient descent to achieve the purpose of classifying the data.

Decision Tree

A decision tree algorithm is a method of building a model based on the characteristics of data using a tree structure [45]. A decision tree is usually composed of nodes and directed edges. The process of constructing decision trees usually includes feature selection, tree generation, and pruning. The essence of decision tree learning is to generalize a set of classification rules from the training dataset, representing a mapping relationship between object attributes and object values.

KNN

The KNN algorithm is used in the case where the data and labels are known in the given training set. The characteristics of the input test data are compared with the corresponding features of the training set to find the top *K* dataset most similar in the training set (ie, the most similar *K* instances, or nearest neighbors), and then the most frequently occurring classification among the *K* most similar data is summarized to classify the test data [46].

Random Forest

Random forest is an integrated learning algorithm that integrates multiple decision trees into a single classifier [47]. The random forest algorithm selects different splitting features and training samples to generate a forest of a large number of decision trees. When predicting unknown samples, each tree in the forest is made to make decisions, which improves the accuracy of the prediction compared to a single decision tree. By statistically determining the results of the decision, the classification with the highest number of votes is taken as the final classification result.

Missing Data

For handling missing values in variables, XGBoost adopts an imputation-free approach in which missing values can be directly marked as “missing” in the input and the model can use only the nonmissing samples for creating trees, so that no value imputation operation was carried out. For the other algorithms, missing values were imputed with the average value of the entire population before model building.

Implementation

All experiments were performed with the web-based interactive tool Jupyter notebook under the environment manager Anaconda, and a python3 kernel was used for data processing and modeling analysis. The XGBoost model relied on the “XGBClassifier” package, and the other machine-learning models were respectively dependent on the “LogisticRegression,” “svm,” “DecisionTreeClassifier,” “RandomForestClassifier,” and “KNeighborsClassifier” packages, which can be accessed from the sklearn library in the public Python software [48,49].

Evaluation Criteria

We used a confusion matrix of the classification results to compute the performance indices, as shown in Table 3.

Based on this confusion matrix, we obtained the following indicators to evaluate the performance of our model. Accuracy was calculated as the proportion of the correct number of samples (true positives [TP]; the true category of the sample is positive and the final predicted result is also positive) to the total number of samples, including false negatives (FN; the true category of the sample is positive and the final predicted result is negative), TP, true negatives (TN; the true category of the sample is negative and the final predicted result is also negative), and false positives (FP; the true category of the sample is negative and the final predicted result is positive) using the following formula: TP+TN/TP+FP+TN+FN.

Sensitivity, also called recall, was calculated as the percentage of TP examples that were correctly predicted: TP/TP+FN.

The positive predictive value (PPV), also known as precision, was calculated as the percentage of positive samples that are predicted correctly: TP/TP+FP.

Specificity was calculated as the proportion of TN samples that was correctly predicted: TN/TN+FP.

The negative predictive value (NPV) was calculated as the percentage of the sample predicted correctly as a negative example: TN/TN+FN.

Finally, the F1-score was calculated as a harmonic average of model accuracy and recall according to the following formula: 2 × (precision×recall)/(precision+recall)

We then sorted the samples according to the prediction results of the model, and predicted the samples as positive examples one by one, successively obtaining the FP rate and TP rate, which were plotted as the horizontal and vertical coordinates to obtain the receiver operating characteristic curve (ROC). The area under the ROC value (AUC) was then selected as the main evaluation index. The more realistic meaning is that given a random positive and a negative sample, the probability of a positive sample output by the classifier is greater than that of negative sample output by the classifier. The formula for calculating AUC is as follows:

Where M represents the number of positive samples, N is the number of negative samples, and $rank_i$ is the order of probability from high to low for positive examples. Therefore, a larger AUC value indicates a better classification result of the learner and a better prediction effect of the model.

Table 3. Confusion matrix.

True_label	Predicted_label	
	Negative example (0)	Positive example (1)
Negative example (0)	True negative	False positive
Positive example (1)	False negative	True positive

Results

Model Prediction Performances

After feature processing, a set of 65 feature variables were finally used as the input of the machine-learning algorithms. We conducted model training, verification, and prediction on the divided training set and test set. The prediction accuracy and AUC values of each model are shown in Table 4, and the detailed ROC curves are depicted in Figure 2. The nonlinear ensemble method XGBoost clearly achieved the highest accuracy on the test dataset. As a similar ensemble method, random forest achieved closely competitive performance. Machine-learning methods with nonlinear models (ie, random forest, KNN classifiers, decision trees, SVM) outperformed the traditional linear logistic regression model that has been widely used in most previous risk prediction models. This suggested that sophisticated machine-learning methods help to improve the performance of risk prediction with a sufficiently sized training dataset.

One potential concern would be that patients in the non-CHD group all had a total follow-up period of >3 years, whereas some patients in the CHD group may have had a follow-up period of less than 3 years, which would likely result in an inherent imbalance between the two groups of data. To exclude the possible bias introduced by variation in the total follow-up time, we carried out an additional experiment in which the test sets were divided into two groups: (1) CHD onset within 3 years and total follow-up >3 years (5094 samples), and (2) CHD onset within 3 years and total follow-up ≤3 years (4984 samples). We applied the same derived prediction model on these two test sets separately, which confirmed that the performance of the model was analogous on both sets with similar AUC values (0.9464 for group 1 vs 0.9389 for group 2; Multimedia Appendix 4) and there was no statistically significant difference on the risk score distributions between the two groups (P=.34 Kolmogorov-Smirnov test). This suggest that the inclusion of CHD patients with under a 3-year follow-up time did not introduce observable data bias and the models developed would be reliable in terms of generalization.

Table 4. Prediction scores of models created by different algorithms.

Algorithm/ model	AUC[a]	ACC[b]	F1-score	Sensitivity	PPV[c]	Specificity	NPV[d]
Logistic regression	0.865	0.809	0.785	0.736	0.840	0.874	0.787
Decision tree	0.882	0.827	0.802	0.742	0.873	0.903	0.796
KNN[e]	0.908	0.827	0.808	0.769	0.851	0.879	0.810
SVM[f]	0.915	0.850	0.832	0.782	0.888	0.912	0.824
Random forest	0.938	0.861	0.846	0.812	0.884	0.905	0.843
XGBoost	0.943	0.870	0.855	0.820	0.895	0.914	0.849

[a]AUC: area under the receiver operating curve.

[b]ACC: accuracy.

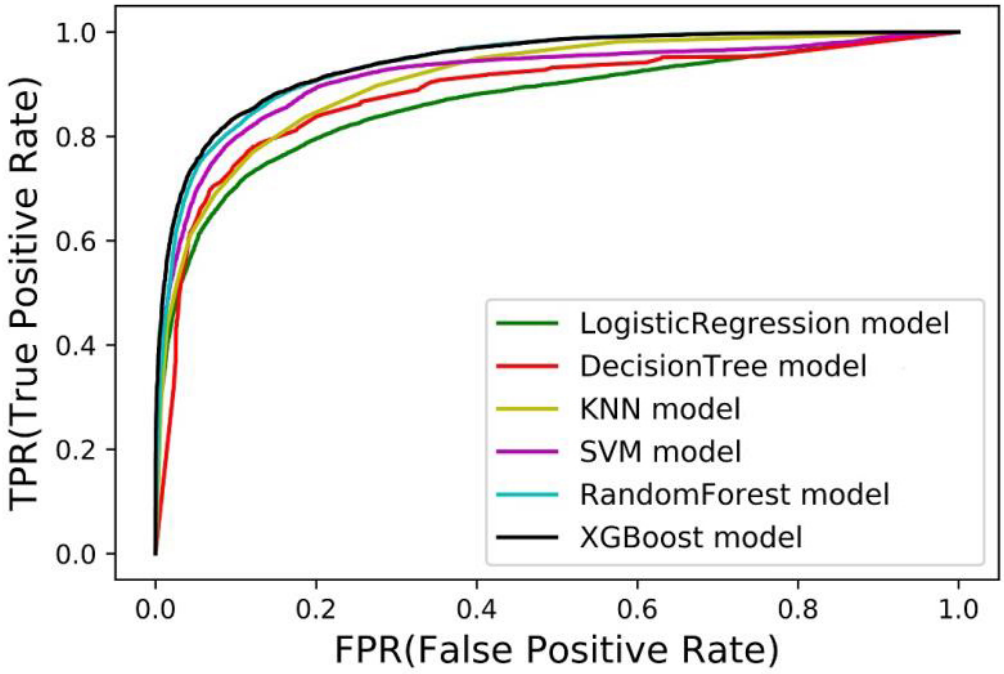
[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

[e]KNN: K-nearest neighbor.

[f]SVM: support vector machine.

Figure 2. The receiver operating characteristic curves of models established by different algorithms. AUC: area under the curve.



Contributions of EHR Feature Variables to Model Prediction

The feature importance of the XGBoost model measures the relative contribution of the feature variables in the process of building the decision trees. Figure 3 depicts the top-ranked features selected by the XGBoost model. In addition to traditional risk factors such as age, systolic blood pressure, and years since hypertension onset, several other features representing the dynamic trends of medical activities also played important roles in risk prediction. For example, an increased frequency of medical activities (in-hospital or out-hospital visits) in the late half or the last half year of the follow-up period would be linked to a higher risk of CHD onset. In addition, the (highest or lowest) blood pressure at the late half of the follow-up period would provide additional information to the risk scores.

To further confirm the contributions of different EHR features on model precision, we performed an experiment in which a series of models were created using an increasing sequence of EHR features and the same XGBoost algorithm, and the performances of these models were tested on the same independent test set. Table 5 summarizes the variation trends of the models with different numbers of features added. Initially, variables that are traditionally used for most risk scales were selected. With only six basic variables, the model reached an AUC of 0.81, which is analogous to the performances of most of the traditional risk scales reported in the literature. Next, diagnosis variables extracted from regular follow-ups, in-hospital, or out-hospital visits were added. Although these symptom data helped to improve the model performance, the effect was quite marginal, which may be attributed to the fact that pre-CHD symptoms are mostly hidden or nonspecific and are often undiagnosed before CHD onset. Finally, variables created by combining multiple EHRs accumulated over time were added. Surprisingly, adding multiple time-point systolic blood pressure values significantly improved the accuracy of the model, suggesting that the long-term variations of blood pressure measurements can be an independent risk factor for CHD prediction. Moreover, variables indicating an increasing trend of medical activities (eg, in-hospital or out-hospital records but without a CHD-related diagnosis or medical examinations) were shown to be correlated with a future risk of CHD onset, which warrants further investigations.

To further analyze the marginal effect of each variable, we performed a univariate trend analysis to describe the relationship between a given variable and the predicted risk probability based on the obtained model, which was visualized with a scatter plot. First, we binned all training samples (including both positive and negative samples) according to the value interval of the studied variable, which was plotted on the x-axis. The corresponding predicted risk probability for each sample was then plotted on the y-axis. Subsequently, a trend curve was plotted showing the averaged risk probability at the given value (or interval) of the studied variable. An example of the marginal effects for four typical variables is depicted in Figure 4. Many variables exhibited highly nonlinear correlations with the overall risk probability scores. This could provide useful insights for CHD prevention through improving risk factor control.

Figure 3. The importance rankings of feature variables for the XGBoost model.

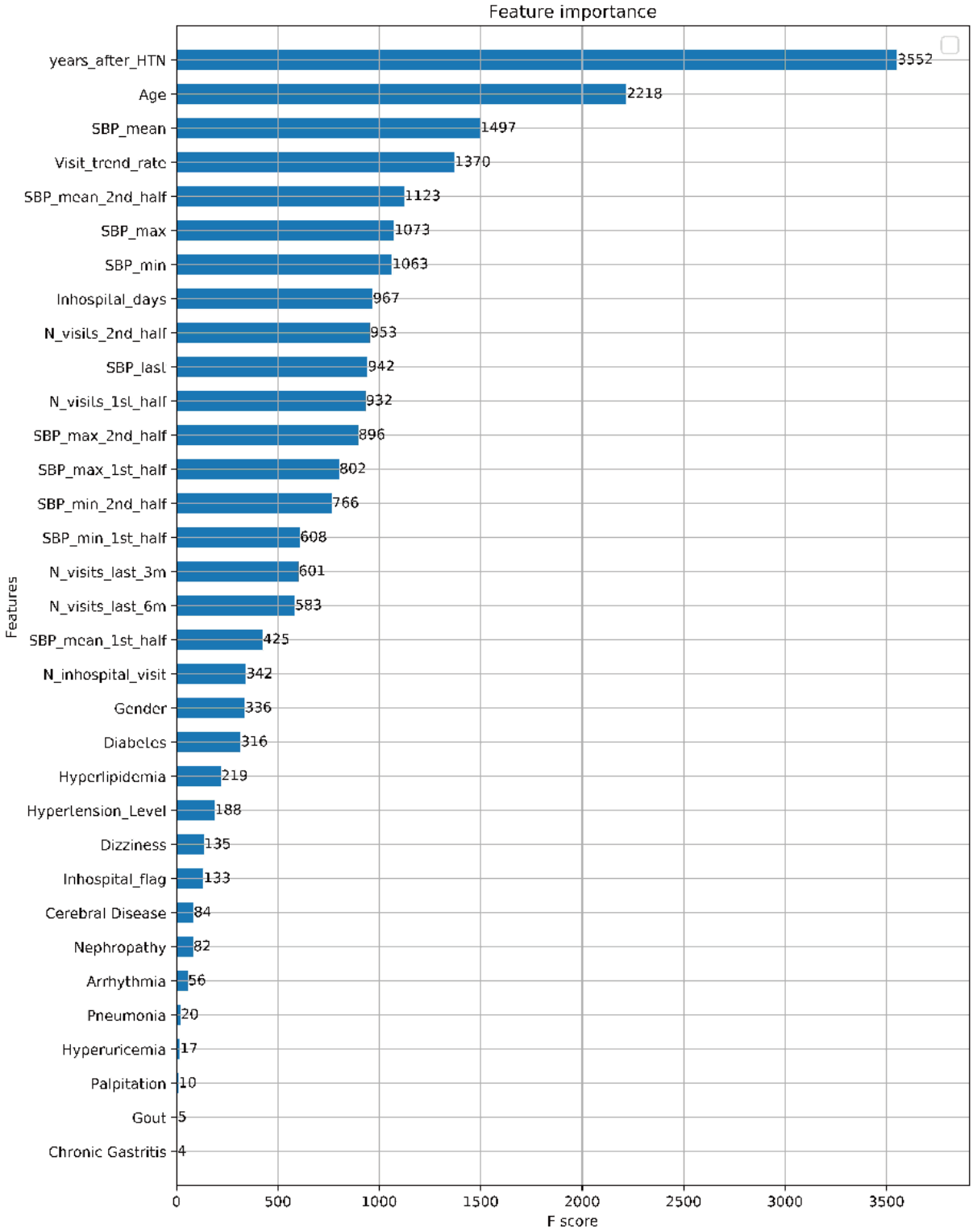
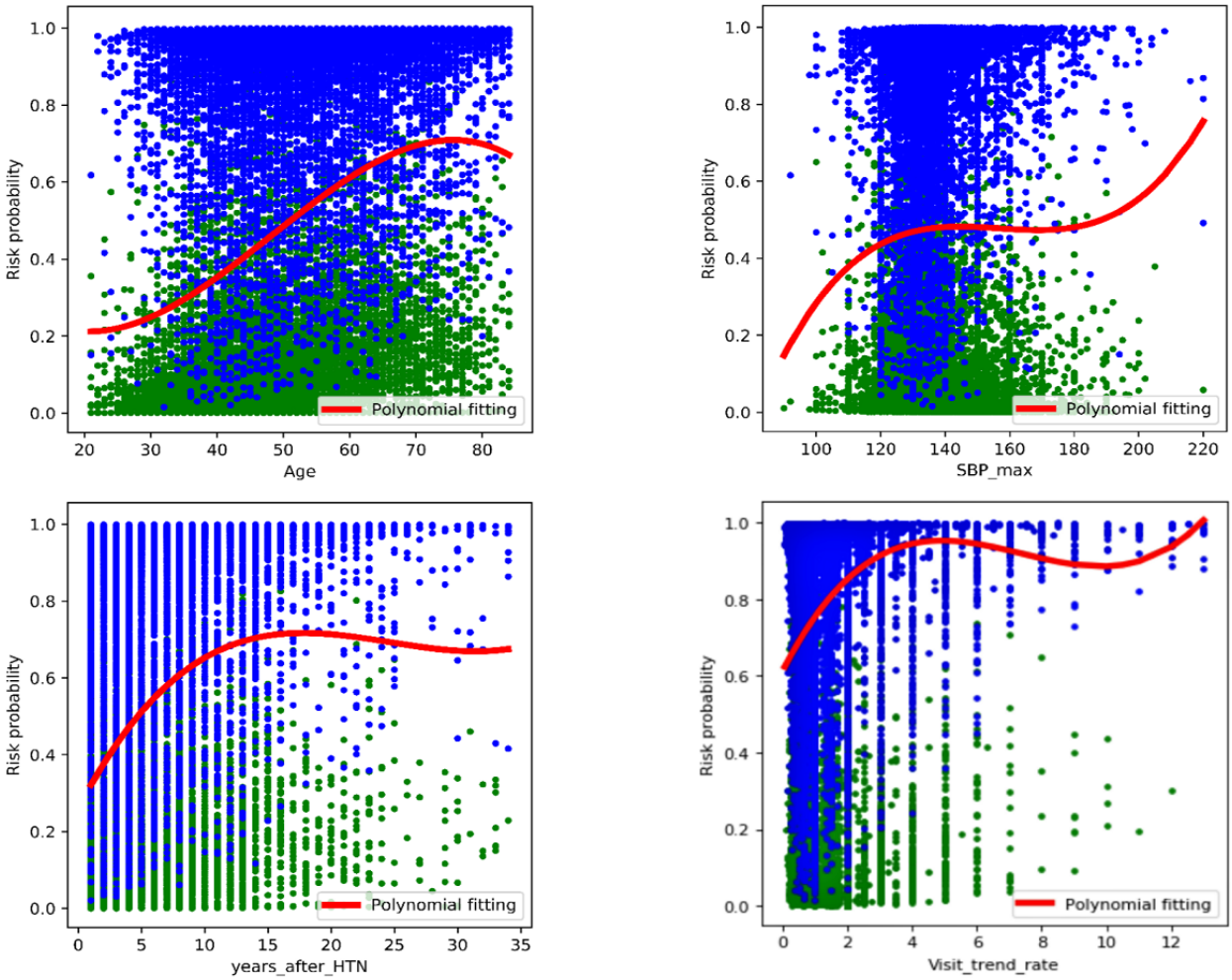


Table 5. Trends of model performance with increasing feature sets.

Variables included in the model	Number of features	AUC[a]	ACC[b]
SBP[c]_last, Age, Gender, Years_After_Hypertension (in the last CHD[d]-free record)	4	0.7547	0.6941
+ Diabetes diagnosis	5	0.7766	0.7090
+ Hyperlipidemia diagnosis	6	0.8111	0.7339
+Inpatient diagnosis flag +Total in-hospital days + Total in-hospital visit number	9	0.8134	0.7341
+ Diagnosed symptoms (eg, hypertension level, cerebral disease, dizziness, nephropathy, gout, hyperuricemia, palpitation)	19	0.8289	0.7460
+ Multipoint SBP statistics (SBP_max, SBP_min, SBP_mean)	22	0.8589	0.7766
+ Dynamic SBP trends (SBP_ min(max.mean)_1st(2nd)_half)	28	0.8752	0.7929
+ Medical activities trends (N_visits_1st_half, N_visits_2nd_half, Visit_trend_ratio)	31	0.9195	0.8350
+ Medical activities trends (N_visits_last_3m[e], N_visits_last_6m[f])	33	0.9427	0.8686

[a]AUC: area under the receiver operating characteristic curve.
[b]ACC: accuracy.
[c]SBP: systolic blood pressure.
[d]CHD: coronary heart disease.
[e]3m: 3 months.
[f]6m: 6 months.

Figure 4. The univariate marginal effects of typical variables on the risk probability scores. The blue dots represent the CHD samples and the green dots represent the non-CHD samples in the training set. The y-axis shows the calculated risk probability scores (0=low risk, 1=high risk). The red curve shows the average risk probability at the given value/interval of the studied variables. CHD: coronary heart disease; SBP: systolic blood pressure; HTN: hypertension.



Impact of Population Size on Model Performance

The creation of most disease prediction models relies on large-scale research cohorts. The size of the research population is one of the important factors that affects the final performance of the created models. To determine the impact of population size on model performance, we carried out an experiment with a series of subpopulations of varying sizes and a fixed number of variables to explore the impact of different data volumes on model performance. The results are depicted in Table 6, demonstrating that the accuracy and reliability of model prediction will be improved with an increase in the size of the research population when the characteristic variables are fixed. However, given adequate variable sets, the model can reach fairly competitive performance (ie, AUC>0.8) even with a small training population size, surpassing the results obtained with a large training population but with limited feature variables (eg, Table 5). This suggested that population size is indeed a very important consideration in building disease risk prediction models but is not an overwhelming limitation.

Table 6. Trends of model accuracy with respect to varying training population size[a].

Training population size (N)	ACC[b]	AUC[c]
200		
Subpopulation 1	0.780	0.850
Subpopulation 2	0.745	0.807
Subpopulation 3	0.740	0.823
Subpopulation 4	0.770	0.840
Subpopulation 5	0.800	0.839
Mean	0.767	0.832
2000		
Subpopulation 1	0.847	0.933
Subpopulation 2	0.838	0.927
Subpopulation 3	0.833	0.921
Subpopulation 4	0.838	0.927
Subpopulation 5	0.835	0.924
Mean	0.838	0.926
20,000		
Subpopulation 1	0.869	0.943
Subpopulation 2	0.868	0.943
Subpopulation 3	0.869	0.943
Subpopulation 4	0.868	0.942
Subpopulation 5	0.870	0.943
Mean	0.869	0.943

[a]For each size, five subpopulations were created and the results were averaged.

[b]ACC: accuracy.

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

Comparison With Traditional Statistical Models

Risk scales obtained by statistical analyses of relatively large samples have long been used in the prevention and screening of the high cardiovascular risk population. Several CHD risk scales have been proposed and widely adopted, such as Framingham risk scales. Therefore, it is also necessary to compare the performance of risk models obtained by machine-learning methods with these traditional risk scales. However, most existing risk scales for CVDs included lifestyle factors and blood test or medical imaging examinations that are not included in routine health checks or chronic disease follow-ups, making it hard to achieve direct comparison with EHR-based studies. In this study, we screened the cohort database to identify a subset of 536 patients (498 with CHD onset and 38 with no CHD onset) with sufficient lifestyle and blood test information required for comparison with the major existing CHD risk scales. These patients were assigned to the test dataset in the first step of our model-building process. We applied the developed XGBoost model as well as the traditional risk scales for these patients, and compared their prediction performance based on the AUC value as the evaluation metric. We should emphasize that due to the low availability in the overall population, some of the features used in the risk scales (such as smoking, diastolic blood pressure, low-density lipoprotein cholesterol, and high-density lipoprotein cholesterol) were not included in the XGBoost risk model. The following three popular risk scales were used for comparison.

The Framingham 10-Years CHD Risk Scale

Proposed by the Framingham Heart Study team in 1998, the Framingham 10-years CHD risk scale is now recognized as an effective tool worldwide to predict the risk and make appropriate preventive management decisions for future CHD onset at the individual level. The age range of the study population is between 30 and 74 years, and the main predictors of this simplified model include gender, age, diabetes, smoking, stratification of blood pressure (systolic and diastolic), and stratification of total cholesterol and high-density lipoprotein cholesterol [50]. It should be noted that the 10-year risk scale was designed for predicting long-term risks, which is somehow divergent from the goal of the present study. However, given that it is one of the most frequently used risk scales, we included the results for reference.

The Framingham 2-Years CHD Risk Scale

Proposed by the Framingham Heart Study team in 2000, the CHD 2-year risk score was developed based on the original

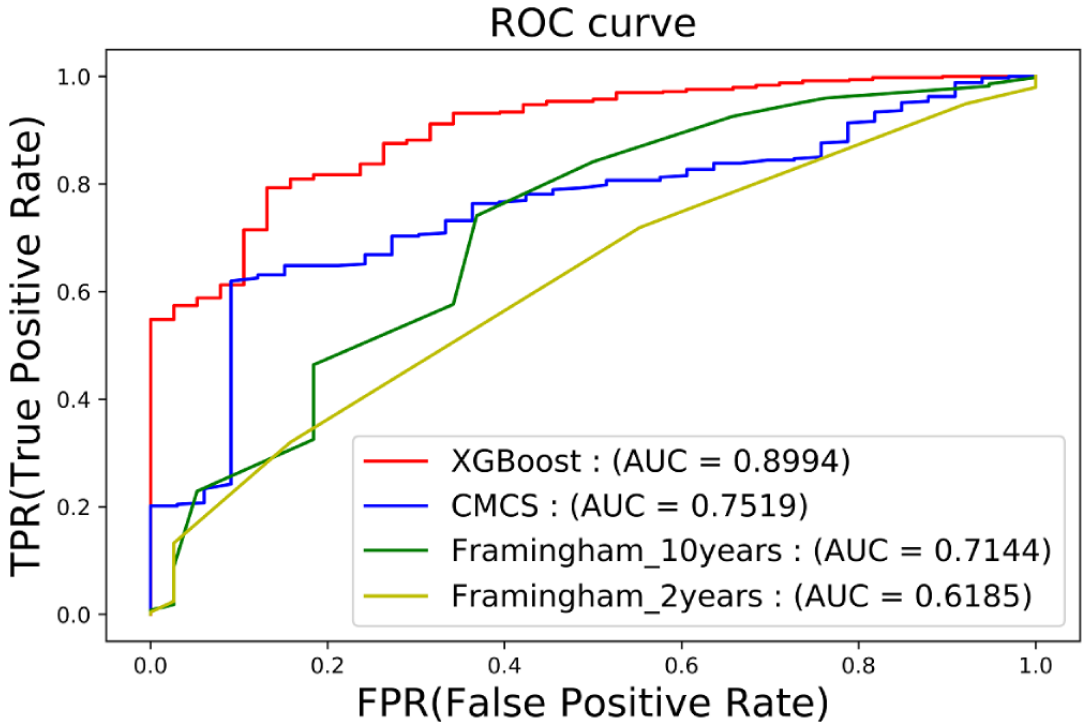
10-year model taking into account updated research results, further deepening and expanding models that predict the risk of recurrent or subsequent CHD events in people with a history of CHD or CVD. The age range of the model population is between 35 and 74 years. The main predictors of this simplified model include gender, age, diabetes, smoking, stratification of blood pressure (systolic), and stratification of total cholesterol and high-density lipoprotein cholesterol [4].

The China Multiprovincial Cohort Study Risk Scale

In 2003, based on a cohort of individuals aged 35 to 64 years living in 11 provinces and cities of China, a risk model for CVD in the Chinese population was established. This model used a prospective cohort study method to calculate the risk factors and the incidence of CVD based on predictive models. The main predictors of this simplified model include gender, age, diabetes, smoking, stratification of blood pressure (systolic), and stratification of total cholesterol and high-density lipoprotein cholesterol [51,52].

Figure 5 shows the ROC curves achieved by the XGBoost model and traditional risk scales. The prediction model established by the XGBoost algorithm showed the best classification performance, with the AUC value reaching 0.8994, followed by the Chinese Multiprovincial Cohort Study queue model, with an AUC value of 0.7519. The prediction accuracy of the Framingham 10-year risk prediction model and 2-year risk prediction model was slightly lower, with AUC values of 0.7144 and 0.6185, respectively. Therefore, our model based on big data and machine-learning algorithms has a better classification effect, higher prediction accuracy, and better performance than traditional statistical models. Moreover, compared with traditional risk scales, our EHR-based model does not require additional medical examinations, which can reduce the patient burden and is beneficial for large-scale population screening.

Figure 5. Comparison of the machine learning-based model and traditional risk scales on the same dataset.



Discussion

We established a high-precision CHD prediction model through EHR big data and machine-learning techniques, and evaluated the effects of different modeling methods, the impact of feature variables, and the dataset size on the model performance. Unlike previous EHR-based studies, our model achieved high prediction accuracy (AUC=0.943) in predicting 3-year CHD onset with the independent test dataset. Further comparison analyses showed that nonlinear models outperform linear models, which was supported by the univariate marginal effect analysis showing that many feature variables had strong nonlinear effects on risk predictions.

We also demonstrated that the construction of secondary feature variables played an important role in the performances of model building. Specifically, we discovered that using time-dependent features obtained from multiple records, including both statistical variables and changing-trend variables, helped to improve the performance rather than using only static features. Moreover, with proper feather variable choices, the prediction model can achieve fairly sufficient precision even when the training sample size is small (compared with datasets from a large population but very few features). This explains the large gap of our models compared with previous EHR-based models.

In summary, our study demonstrated that accurate prediction of 3-year CHD onset risk is possible for a large group of patients with hypertension solely based on EHR data collected during routing follow-up visits for chronic diseases with in-hospital and out-hospital diagnostic records. Using an independent test dataset, we verified that EHR-based models can achieve better

risk prediction performance than traditional risk scales. Compared with traditional risk scales, the EHR-based model does not involve additional medical examinations, which reduces the patient burden and is beneficial for large-scale population screening. Moreover, compared with traditional patient cohort studies, EHR-based studies are far easier to conduct with respect to data acquisition and facilitate investigating many variables in a batch simultaneously. Our results indicate that long-term accumulation of EHR big data through centralized platforms, especially the multiple time-point changes of patient health status, provides very important information for the prediction and early prevention of chronic diseases. Further investigations are needed to explore the power of accumulated historical data.

The major limitation of our study is that we used anonymized historical EHR data, which had a high missing rate. Some known potential risk factors such as diastolic blood pressure, BMI, and blood test indicators were not considered as important factors in the modeling process because of the large proportion of data missing in the population. The missing data also affected the acquisition of outcome status for each patient. The CHD onset label can be imprecise if the patient did not receive a hospital diagnosis during the study period and within the regional hospital system. This is a defect compared with traditional cohort studies. However, the impact of missing information is equal for both the positive and negative groups so that no significant biases are likely to be introduced through missing data. Compared with the benefits obtained by the enlarged population and the abundance of clinical features, the increased noise in the data is considered to be acceptable.

Acknowledgments

This research is supported by the Strategic Priority CAS project (XDB38000000), Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Major R&D Project of Guangdong (2017B030308007), and Shenzhen Science and Technology Research Funding (20170502165510880).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of fields for data collection from electronic health records.
[DOCX File , 14 KB - medinform_v8i7e17257_app1.docx]

Multimedia Appendix 2
Distribution of predicted scores in the stroke group (N=10,183).
[PNG File , 383 KB - medinform_v8i7e17257_app2.png]

Multimedia Appendix 3
Distribution of coronary heart disease (CHD)-free time in the CHD group (0-3 years, N=20,156).
[PNG File , 361 KB - medinform_v8i7e17257_app3.png]

Multimedia Appendix 4
Receiver operating characteristic (ROC) curves according to coronary heart disease (CHD) onset and follow-up times.
[PNG File , 605 KB - medinform_v8i7e17257_app4.png]

References

1. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. J Am Coll Cardiol 2017 Jul 04;70(1):1-25 [FREE Full text] [doi: 10.1016/j.jacc.2017.04.052] [Medline: 28527533]
2. World Health Organization. 2017 May. Cardiovascular diseases (CVDs): Key Facts URL: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2017-05-17]
3. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol 1976 Jul;38(1):46-51. [doi: 10.1016/0002-9149(76)90061-8] [Medline: 132862]
4. D'Agostino RB, Russell MW, Huse DM, Ellison R, Silbershatz H, Wilson PW, et al. Primary and subsequent coronary risk appraisal: new results from the Framingham study. Am Heart J 2000 Feb;139(2 Pt 1):272-281. [doi: 10.1067/mhj.2000.96469] [Medline: 10650300]
5. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008 Feb 12;117(6):743-753. [doi: 10.1161/CIRCULATIONAHA.107.699579] [Medline: 18212285]

6. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2014 Jul 01;63(25 Pt B):2935-2959 [FREE Full text] [doi: 10.1016/j.jacc.2013.11.005] [Medline: 24239921]
7. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ 2007 Jul 21;335(7611):136 [FREE Full text] [doi: 10.1136/bmj.39261.471806.55] [Medline: 17615182]
8. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008 Jun 28;336(7659):1475-1482 [FREE Full text] [doi: 10.1136/bmj.39609.449676.25] [Medline: 18573856]
9. Conroy R, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 2003 Jun;24(11):987-1003. [doi: 10.1016/s0195-668x(03)00114-3] [Medline: 12788299]
10. Woodward M, Tunstall-Pedoe H, Peters SA. Graphics and statistics for cardiology: clinical prediction rules. Heart 2017 Apr;103(7):538-545 [FREE Full text] [doi: 10.1136/heartjnl-2016-310210] [Medline: 28179372]
11. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016 May 16;353:i2416 [FREE Full text] [doi: 10.1136/bmj.i2416] [Medline: 27184143]
12. Karmali KN, Persell SD, Perel P, Lloyd-Jones DM, Berendsen MA, Huffman MD. Risk scoring for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev 2017 Mar 14;3:CD006887 [FREE Full text] [doi: 10.1002/14651858.CD006887.pub4] [Medline: 28290160]
13. Ahmad S, Moorthy MV, Demler OV, Hu FB, Ridker PM, Chasman DI, et al. Assessment of Risk Factors and Biomarkers Associated With Risk of Cardiovascular Disease Among Women Consuming a Mediterranean Diet. JAMA Netw Open 2018 Dec 07;1(8):e185708 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.5708] [Medline: 30646282]
14. Estruch R, Ros E, Salas-Salvadó J, Covas MI, Corella D, Arós F, PREDIMED Study Investigators. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts. N Engl J Med 2018 Jun 21;378(25):e34. [doi: 10.1056/NEJMoa1800389] [Medline: 29897866]
15. Domínguez F, Fuster V, Fernández-Alvira JM, Fernández-Friera L, López-Melgar B, Blanco-Rojo R, et al. Association of Sleep Duration and Quality With Subclinical Atherosclerosis. J Am Coll Cardiol 2019 Jan 22;73(2):134-144 [FREE Full text] [doi: 10.1016/j.jacc.2018.10.060] [Medline: 30654884]
16. Shrivastava AK, Singh HV, Raizada A, Singh SK. C-reactive protein, inflammation and coronary heart disease. Egypt Heart J 2015 Jun;67(2):89-97. [doi: 10.1016/j.ehj.2014.11.005]
17. Parrinello CM, Lutsey PL, Ballantyne CM, Folsom AR, Pankow JS, Selvin E. Six-year change in high-sensitivity C-reactive protein and risk of diabetes, cardiovascular disease, and mortality. Am Heart J 2015 Aug;170(2):380-389 [FREE Full text] [doi: 10.1016/j.ahj.2015.04.017] [Medline: 26299237]
18. Matsushita K, Coresh J, Sang Y, Chalmers J, Fox C, Guallar E, CKD Prognosis Consortium. Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: a collaborative meta-analysis of individual participant data. Lancet Diabetes Endocrinol 2015 Jul;3(7):514-525 [FREE Full text] [doi: 10.1016/S2213-8587(15)00040-6] [Medline: 26028594]
19. Bent RE, Wheeler MT, Hadley D, Knowles JW, Pavlovic A, Finocchiaro G, et al. Systematic Comparison of Digital Electrocardiograms From Healthy Athletes and Patients With Hypertrophic Cardiomyopathy. J Am Coll Cardiol 2015 Jun 09;65(22):2462-2463 [FREE Full text] [doi: 10.1016/j.jacc.2015.03.559] [Medline: 26046742]
20. Hagnäs MJ, Lakka TA, Kurl S, Rauramaa R, Mäkikallio TH, Savonen K, et al. Cardiorespiratory fitness and exercise-induced ST segment depression in assessing the risk of sudden cardiac death in men. Heart 2017 Mar;103(5):383-389. [doi: 10.1136/heartjnl-2015-309217] [Medline: 27604814]
21. Ackerman MJ, Zipes DP, Kovacs RJ, Maron BJ. Eligibility and Disqualification Recommendations for Competitive Athletes With Cardiovascular Abnormalities: Task Force 10: The Cardiac Channelopathies: A Scientific Statement From the American Heart Association and American College of Cardiology. J Am Coll Cardiol 2015 Dec 01;66(21):2424-2428 [FREE Full text] [doi: 10.1016/j.jacc.2015.09.042] [Medline: 26542662]
22. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. Nat Rev Cardiol 2015 Nov;12(11):627-642. [doi: 10.1038/nrcardio.2015.152] [Medline: 26461967]
23. Lee JJ, Pedley A, Hoffmann U, Massaro JM, Fox CS. Association of Changes in Abdominal Fat Quantity and Quality With Incident Cardiovascular Disease Risk Factors. J Am Coll Cardiol 2016 Oct 04;68(14):1509-1521 [FREE Full text] [doi: 10.1016/j.jacc.2016.06.067] [Medline: 27687192]
24. Hecht HS, Cronin P, Blaha MJ, Budoff MJ, Kazerooni EA, Narula J, et al. 2016 SCCT/STR guidelines for coronary artery calcium scoring of noncontrast noncardiac chest CT scans: A report of the Society of Cardiovascular Computed Tomography and Society of Thoracic Radiology. J Cardiovasc Comput Tomogr 2017;11(1):74-84. [doi: 10.1016/j.jcct.2016.11.003] [Medline: 27916431]

25. Bacharova L, Chen H, Estes EH, Mateasik A, Bluemke DA, Lima JA, et al. Determinants of discrepancies in detection and comparison of the prognostic significance of left ventricular hypertrophy by electrocardiogram and cardiac magnetic resonance imaging. Am J Cardiol 2015 Feb 15;115(4):515-522 [FREE Full text] [doi: 10.1016/j.amjcard.2014.11.037] [Medline: 25542394]
26. Buchanan C, Mohammed A, Cox E, Köhler K, Canaud B, Taal MW, et al. Intradialytic Cardiac Magnetic Resonance Imaging to Assess Cardiovascular Responses in a Short-Term Trial of Hemodiafiltration and Hemodialysis. J Am Soc Nephrol 2017 Apr;28(4):1269-1277 [FREE Full text] [doi: 10.1681/ASN.2016060686] [Medline: 28122851]
27. Kubo T, Shinke T, Okamura T, Hibi K, Nakazawa G, Morino Y, OPINION Investigators. Optical frequency domain imaging vs. intravascular ultrasound in percutaneous coronary intervention (OPINION trial): one-year angiographic and clinical results. Eur Heart J 2017 Nov 07;38(42):3139-3147 [FREE Full text] [doi: 10.1093/eurheartj/ehx351] [Medline: 29121226]
28. Arsenault BJ, Lachance D, Lemieux I, Alméras N, Tremblay A, Bouchard C, et al. Visceral adipose tissue accumulation, cardiorespiratory fitness, and features of the metabolic syndrome. Arch Intern Med 2007 Jul 23;167(14):1518-1525. [doi: 10.1001/archinte.167.14.1518] [Medline: 17646606]
29. de Franciscis S, Metzinger L, Serra R. The Discovery of Novel Genomic, Transcriptomic, and Proteomic Biomarkers in Cardiovascular and Peripheral Vascular Disease: The State of the Art. Biomed Res Int 2016;2016:7829174. [doi: 10.1155/2016/7829174] [Medline: 27298828]
30. Ngo D, Sinha S, Shen D, Kuhn EW, Keyes MJ, Shi X, et al. Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease. Circulation 2016 Jul 26;134(4):270-285 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.116.021803] [Medline: 27444932]
31. Ferguson JF, Allayee H, Gerszten RE, Ideraabdullah F, Kris-Etherton PM, Ordovás JM, American Heart Association Council on Functional GenomicsTranslational Biology, Council on EpidemiologyPrevention,Stroke Council. Nutrigenomics, the Microbiome, and Gene-Environment Interactions: New Directions in Cardiovascular Disease Research, Prevention, and Treatment: A Scientific Statement From the American Heart Association. Circ Cardiovasc Genet 2016 Jun;9(3):291-313. [doi: 10.1161/HCG.0000000000000030] [Medline: 27095829]
32. Oikonomou E, Lazaros G, Georgiopoulos G, Christoforatou E, Papamikroulis GA, Vogiatzi G, et al. Environment and cardiovascular disease: rationale of the Corinthia study. Hellenic J Cardiol 2016;57(3):194-197 [FREE Full text] [doi: 10.1016/j.hjc.2016.06.001] [Medline: 27451913]
33. Bhatnagar A. Environmental Determinants of Cardiovascular Disease. Circ Res 2017 Jul 07;121(2):162-180 [FREE Full text] [doi: 10.1161/CIRCRESAHA.117.306458] [Medline: 28684622]
34. Pearce C, Bainbridge M. A personally controlled electronic health record for Australia. J Am Med Inform Assoc 2014;21(4):707-713 [FREE Full text] [doi: 10.1136/amiajnl-2013-002068] [Medline: 24650635]
35. Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah AD, Denaxas S, et al. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. Lancet 2014 May 31;383(9932):1899-1911 [FREE Full text] [doi: 10.1016/S0140-6736(14)60685-1] [Medline: 24881994]
36. Bandyopadhyay S, Wolfson J, Vock DM, Vazquez-Benitez G, Adomavicius G, Elidrisi M, et al. Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. Data Min Knowl Disc 2014 Oct 4;29(4):1033-1069. [doi: 10.1007/s10618-014-0386-6]
37. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
38. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12(4):e0174944 [FREE Full text] [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]
39. Bell S, Daskalopoulou M, Rapsomaniki E, George J, Britton A, Bobak M, et al. Association between clinically recorded alcohol consumption and initial presentation of 12 cardiovascular diseases: population based cohort study using linked health records. BMJ 2017 Mar 22;356:j909 [FREE Full text] [doi: 10.1136/bmj.j909] [Medline: 28331015]
40. World Health Organization. 2004. ICD-10: international statistical classification of diseases and related health problems: tenth revision, 2nd edition URL: https://apps.who.int/iris/handle/10665/42980 [accessed 2020-05-29]
41. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. 2016 Aug Presented at: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
42. Cortes C, Vapnik V. Support-vector networks. Machine Learn 1995 Sep;20:273-297 [FREE Full text]
43. Park E, Chang H, Nam HS. Use of Machine Learning Classifiers and Sensor Data to Detect Neurological Deficit in Stroke Patients. J Med Internet Res 2017 Apr 18;19(4):e120 [FREE Full text] [doi: 10.2196/jmir.7092] [Medline: 28420599]
44. Hosmer JDW, Lemeshow S, Sturdivant RX. Applied logistic regression. Canada: John Wiley & Sons; 2013.
45. Barros RC, Basgalupp MP, de Carvalho ACPLF, Freitas AA. A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Trans Syst Man Cybern C 2012 May;42(3):291-312. [doi: 10.1109/tsmcc.2011.2157494]

46. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In: Meersman R, Tari Z, Schmidt DC, editors. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science Vol. 2888. Berlin, Heidelberg: Springer; 2003:986-996.
47. Breiman L. Random forests. Machine Learn 2001;45(1):5-32. [doi: 10.1201/9780367816377-11]
48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Machine Learn Res 2011;12:2825-2830.
49. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning softwarexperiences from the scikit-learn project. arXiv preprint 2013 Sep:1309.0238 [FREE Full text]
50. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998 May 12;97(18):1837-1847. [doi: 10.1161/01.cir.97.18.1837] [Medline: 9603539]
51. Wang W, Zhao D, Liu J. Prospective study on the predictive model of cardiovascular disease risk in a Chinese population aged 35-64J. Zhonghua Xin Xue Guan Bing Za Zhi 2003;31(12):902-908. [doi: 10.3760/j:issn:0253-3758.2003.12.006]
52. Liu J, Zhao D, Wang W. Comparison between the results from the Chinese Multi-provincial Cohort Study and those from the Framingham Heart StudyJ. Chinese J Cardiol 2004;32(2):167-172. [doi: 10.3760/j:issn:0253-3758.2004.02.020]

Abbreviations

AUC: area under the receiver operating characteristic curve
CHD: coronary heart disease
CVD: cardiovascular disease
EHR: electronic health record
FN: false negative
FP: false positive
ICD: International Statistical Classification of Diseases and Related Health Problems
KNN: K-nearest neighbor
NPV: negative predictive value
PPV: positive predictive value
ROC: receiver operating characteristic
SVM: support vector machine
TN: true negative
TP: true positive

Edited by G Eysenbach; submitted 29.11.19; peer-reviewed by E Ding, S Veeranki; comments to author 27.01.20; revised version received 09.03.20; accepted 28.03.20; published 06.07.20.

Please cite as:
Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, Fan J, Cheng W, Chen XH, Cai Y
Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation
JMIR Med Inform 2020;8(7):e17257
URL: https://medinform.jmir.org/2020/7/e17257
doi:10.2196/17257
PMID:32628616

Original Paper

Precision Health–Enabled Machine Learning to Identify Need for Wraparound Social Services Using Patient- and Population-Level Data Sets: Algorithm Development and Validation

Suranga N Kasthurirathne[1,2], PhD; Shaun Grannis[1,2], MS, MD; Paul K Halverson[3], DrPH, FACHE; Justin Morea[2,4], MS, MBA, DO; Nir Menachemi[1,3], MPH, PhD; Joshua R Vest[1,3], MPH, PhD

[1]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, United States

[2]School of Medicine, Indiana University, Indianapolis, IN, United States

[3]Richard M Fairbanks School of Public Health, Indiana University, Indianapolis, IN, United States

[4]Eskenazi Health, Indianapolis, IN, United States

Corresponding Author:
Suranga N Kasthurirathne, PhD
Center for Biomedical Informatics
Regenstrief Institute
1101 W 10th Street
Indianapolis, IN, 46202
United States
Phone: 1 3172749000
Email: snkasthu@iu.edu

Abstract

Background: Emerging interest in precision health and the increasing availability of patient- and population-level data sets present considerable potential to enable analytical approaches to identify and mitigate the negative effects of social factors on health. These issues are not satisfactorily addressed in typical medical care encounters, and thus, opportunities to improve health outcomes, reduce costs, and improve coordination of care are not realized. Furthermore, methodological expertise on the use of varied patient- and population-level data sets and machine learning to predict need for supplemental services is limited.

Objective: The objective of this study was to leverage a comprehensive range of clinical, behavioral, social risk, and social determinants of health factors in order to develop decision models capable of identifying patients in need of various wraparound social services.

Methods: We used comprehensive patient- and population-level data sets to build decision models capable of predicting need for behavioral health, dietitian, social work, or other social service referrals within a safety-net health system using area under the receiver operating characteristic curve (AUROC), sensitivity, precision, F1 score, and specificity. We also evaluated the value of population-level social determinants of health data sets in improving machine learning performance of the models.

Results: Decision models for each wraparound service demonstrated performance measures ranging between 59.2%% and 99.3%. These results were statistically superior to the performance measures demonstrated by our previous models which used a limited data set and whose performance measures ranged from 38.2% to 88.3% (behavioural health: F1 score $P<.001$, AUROC $P=.01$; social work: F1 score $P<.001$, AUROC $P=.03$; dietitian: F1 score $P=.001$, AUROC $P=.001$; other: F1 score $P=.01$, AUROC $P=.02$); however, inclusion of additional population-level social determinants of health did not contribute to any performance improvements (behavioural health: F1 score $P=.08$, AUROC $P=.09$; social work: F1 score $P=.16$, AUROC $P=.09$; dietitian: F1 score $P=.08$, AUROC $P=.14$; other: F1 score $P=.33$, AUROC $P=.21$) in predicting the need for referral in our population of vulnerable patients seeking care at a safety-net provider.

Conclusions: Precision health–enabled decision models that leverage a wide range of patient- and population-level data sets and advanced machine learning methods are capable of predicting need for various wraparound social services with good performance.

(JMIR Med Inform 2020;8(7):e16129) doi:10.2196/16129

KEYWORDS

social determinants of health; supervised machine learning; delivery of health care; integrated; wraparound social services

Introduction

Background

The combination of precision health [1] and population health initiatives in the United States have raised awareness about how clinical, behavioral, social risk, and social determinants of health factors influence an individual's use of medical services and their overall health and well-being [2]. Large-scale adoption of health information systems [3], increased use of interoperable health information exchange, and the availability of socioeconomic data sets have led to unprecedented and ever increasing accessibility to various patient- and population-level data sources. The availability of these data sets, together with a focus on mitigating patient social factors and uptake of machine learning solutions for health care present considerable potential for predictive modeling in support of risk prediction and intervention allocation [4,5]. This is particularly significant for wraparound services that can enhance primary care by utilizing providers who are trained in behavioral health, social work, nutritional counseling, patient navigation, health education, and medical legal partnerships in order to mitigate the effects of social risk and to address social needs [6].

Wraparound services focus on the socioeconomic, behavioral, and financial factors that typical medical care encounters cannot address satisfactorily [7,8], and when used, can result in improved health care outcomes, reduced costs [6,9], and better coordination of care. As such, these services are of significant importance to health care organizations that are incentivized by United States reimbursement policies to mitigate the effects of social issues that influence poor health outcomes and unnecessary utilization of costly services [10].

Previous Work

In a previous study [11], we integrated patient-level clinical, demographic, and visit data with population-level social determinants of health measures to develop decision models that predicted patient need for behavioral health, dietitian, social work, or other wraparound service referrals. We also compared the performance of models built with and without population-level social determinants of health indicators. These models achieved reasonable performance with area under the receiver operating characteristic curve values between 70% and 78%, and sensitivity, specificity, and accuracy values ranging between 50% and 77%. We integrated these models into nine federally qualified health center sites operated by Eskenazi Health, a county-owned safety-net provider located in Indianapolis, Indiana. A subsequent trial identified increased rates of referral when predicted-need scores were shared with primary care end users [12]. Nevertheless, there were several limitations in our previous study such as limited patient-level measures, a level of aggregated data that was too coarse, poor optimization, lack of consideration of data temporality, and limited generalizability.

Our previous models included a wide range of patient-level clinical, behavioral, and encounter-based data elements as well as population-level social determinants of health measures; however, the models might have performed better with the inclusion of additional data elements such as medication data, insurance information, narcotics or substance abuse data, mental and behavioral disorders information inferred from diagnostic data, and patient-level social risk factors extracted from diagnostic data using ICD-10 classification codes [13].

Our previous use of population-level social determinants of health factors measured at the zip-code level did not contribute to any statistically significant performance improvements. A wider range of measures of social determinants of health captured at smaller geographic areas might have yielded more discriminative power and have led to significant performance improvements.

We used Youden J-index [14] which optimizes sensitivity and specificity to determine optimal cutoff thresholds; however, this resulted in poor precision (positive predictive values that ranged between 15% and 50%). Given the importance of optimizing precision, which represents a model's ability to return only relevant instances, alternate optimization techniques should be used.

Our previous models included all data captured during the period under study, and not exclusively data elements that occurred prior to the outcome of interest. Failing to omit data elements that occurred after the outcomes of interest may have influenced the performance of these decision models [15].

We developed our previous approach using data that was extracted from a homegrown electronic health record system [16]. This limited its ability to be replicated across other settings that could support other widely used commercial electronic health record systems. Since our previous study, Eskenazi Health has transitioned to a commercial electronic health record system enabling us to adapt our solution to be vendor neutral and applicable to any electronic health record system.

Objective

This study addressed the aforementioned limitations by using additional patient- and population-level data elements as well as more advanced analytical methods to develop decision models to identify patients in need of referral to providers that can address social factors. We evaluated the contribution of these enhancements by recreating the original models that had been developed during the previous study (phase 1) and comparing their performance to that of new models developed during this study (phase 2). Furthermore, during each phase, we evaluated the contribution of small-area population-level social determinants of health measures to improving model performance.

Methods

Patient Sample

We included adults (18 years of age or older) with at least one outpatient visit at Eskenazi Health between October 1, 2016 and May 1, 2018.

Data Extraction

Primary data sources for the patient cohort were Eskenazi Health's Epic electronic health record system and the statewide health information exchange data repository known as the Indiana Network for Patient Care [17], which provided out-of-network encounter data from hospitals, laboratory systems, long-term care facilities, and federally qualified health centers across the state. These data were supplemented with population-level social determinants of health measures derived from the US Census Bureau, the Marion County Public Health Department vital statistics system, and various community health surveys.

Feature Extraction

To recreate the models developed during phase 1, we extracted a subset of features that had been used to train the original models [11]. We also extracted additional features for phase 2 enhancements. Table 1 presents an outline of the feature sets for each phase of model development.

Table 1. Comparison of the patient- and population-level data sets that were used for each phase.

Feature type	Phase 1	Added in phase 2
Demographics	Age, ethnicity, and gender	Insurance (Medicare, Medicaid, self-pay)
Weight and nutrition	None	BMI, hemoglobin A_{1c}
Encounter frequency	Outpatient visits, emergency department encounters, and inpatient admissions	None
Chronic conditions	20 most common chronic conditions [18]	None
Addictions and narcotics use	Tobacco and opioid use	Alcohol abuse, opioid overdose, use disorders
Medications	None	145 categories of medication (categorized by therapeutic and pharmaceutical codes) [19]
Patient-level social risk	None	12 patient-level measures [20]
Population-level social determinants of health	48 social determinants of health measures [11]	60 social determinants of health measures [20]

Preparation of the Gold Standard

We sought to predict the need for referrals to behavioral health services, dietitian counseling, social work services, and all other wraparound services, which included respiratory therapy, financial planning, medical-legal partnership assistance, patient navigation, and pharmacist consultations. We used billing, encounter, and scheduling data extracted from the Indiana Network for Patient Care and Eskenazi Health to identify patients who had been referred to supplementary services between October 1, 2016 and May 1, 2018. We assumed that a patient with a referral had been in need of that referral even if the patient subsequently canceled or failed to keep the appointment.

Data Vector Preparation

We prepared two data vectors for each wraparound service for phase 1 modeling—a clinical data vector consisting of only patient-level data elements and a master data vector consisting of both patient- and population-level elements. Next, we created two more data vectors for each wraparound service for phase 2 data—a clinical data vector consisting of only patient-level data elements and a master data vector consisting of both patient- and population-level elements. For each patient, we included only data for events that had occurred at least 24 hours prior to the final outcome of interest. Features such as age (discrete by whole years); weight- or nutrition-based (categorical); gender (categorical); ethnicity (categorical); encounter frequency (number of each type per patient); and addictions or use of narcotics, chronic conditions, medications, and patient-level social risk (binary indicating presence or absence).

Population-level social determinants of health measures were categorized into three groups—socioeconomic status, disease prevalence, and other miscellaneous factors (such as data on calls made by those who were seeking public assistance). Measures that were reported from across 1150 census tracts were used to calculate *z* scores (a numerical measurement relating a given value to the mean in a group of values) for each of the three categories. The *z* scores were grouped into clusters using the *k*-means algorithm [21] and the elbow method [22].

As requested by dietitians who consulted on our efforts, for dietitian referrals, prediction of need was restricted to a subset of patients with specific risk conditions (Multimedia Appendix 1). Thus, data vectors for dietitian referrals included only patients with one or more of these conditions, which were identified by ICD-10 classification codes.

Machine Learning Process for Phase 1 Models

We randomly split each data vector into groups of 80% (training and validation data set) and 20% (test set). We replicated the same processes that were used during phase 1 [11] to recreate a new set of models to be used for comparison.

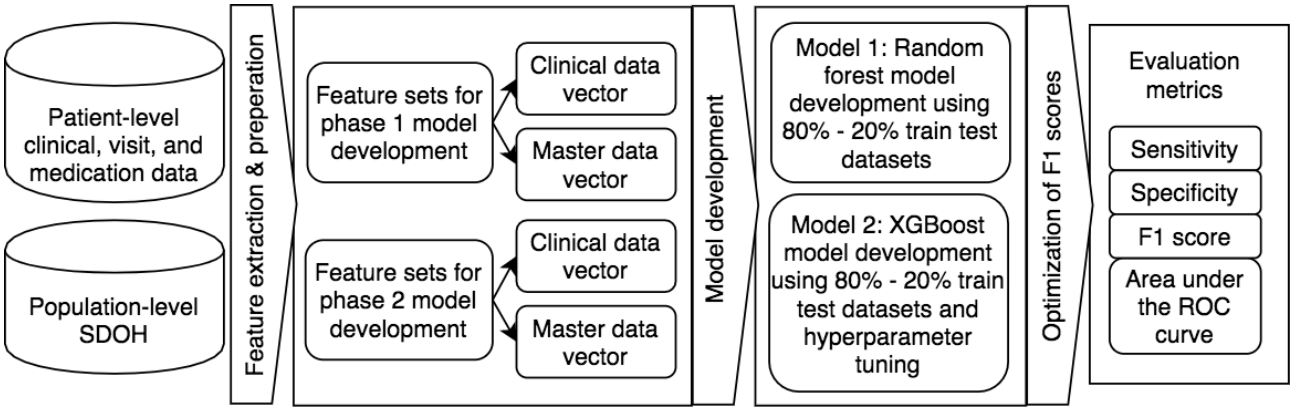
Machine Learning Process for Phase 2 Models

We split each data vector into random groups of 80% (training and validation data set) and 20% (test set). We applied randomized lasso-based [23] feature selection to the 80% training and validation data set to identify the most relevant features for each outcome of interest. We used machine learning in Python (version 3.6.1; scikit-learn library, version 0.21.0) [24] to build extreme gradient boosting [25] classification models to predict the need for referrals. The extreme gradient boosting algorithm is an implementation of gradient boosted decision trees [26] designed for speed and performance. It has demonstrated a strong track record of outperforming other decision trees and other classification algorithms in machine learning competitions [27]. The extreme gradient boosting algorithm consisted of multiple parameters, each of which could affect model performance. Thus, we decided to perform hyperparameter tuning on the training and validation data set using randomized search and 10-fold cross-validation. Decision model parameters that were modified as part of the hyperparameter tuning process are listed in Multimedia Appendix 2. The best performing models, parameterized using hyperparameter tuning, were applied to the test data sets.

Analysis

We assessed the performance of each decision model using the test set. For each record in the test set, each decision model produced a binary outcome (referral needed or referral not needed) and a probability score. We used these scores to calculate area under the receiver operating characteristic curve (AUROC), sensitivity, precision, F1 score, and specificity for each model. These measures were calculated using thresholds that optimized sensitivity and precision scores. We also calculated 95% confidence intervals for each measure using bootstrap methods [28]. *P* values were calculated using guidelines presented by Altman and Bland [29]. *P* values<.05 were deemed statistically significant. For the models trained during each phase, we evaluated the contribution of population-level measures by comparing the performance of models trained using master (with population-level measures) vector models to the performance of clinical (without population-level measures) vector models. Next, we evaluated the value of the additional data sets and analytical methods that were used to train phase 2 models by comparing their performance to that of models trained in phase 1. Figure 1 presents a flowchart that describes the approach.

Figure 1. The complete study approach from data collection and decision-model building to evaluation of results. ROC: receiver operating characteristic; SDOH: social determinants of health; XGBoost: extreme gradient boosting.



Results

Our patient sample consisted of 72,484 adult patients (Table 2). Of these patients, 15,867 (21.9%) met the dietitian referral criteria. Similar to that of phase 1, our patient population reflected an adult, urban, low-income primary care safety-net population; patients ranged in age from 18 to 107 years and were predominantly female (47,187/72,484, 65.1%). Referral types, which constituted our gold standard reference, were behavioral health (12,162/72,484, 16.8%), social work (4104/72,484, 5.7%), dietitian counseling (4330/15,867, 27.3%), and other services (17,877/72,484, 24.7%).

As with our previous effort, use of population-level social determinants of health measures led to only minimal changes in each performance metric across models trained under phases 1 and 2, and were not statistically significant (behavioural health: F1 score *P*=.08, AUROC *P*=.09; social work: F1 score *P*=.16, AUROC *P*=.09; dietitian: F1 score *P*=.08, AUROC *P*=.14; other: F1 score *P*=.33, AUROC *P*=.21). Thus, we evaluated the contribution of the additional data sets, classification algorithms, and analytical approaches leveraged in phase 2 by comparing clinical vector models developed during phase 1 to those developed during phase 2.

Table 3 presents a comparison of clinical vector model performance for phase 1 and phase 2. Phase 2 models yielded significantly better results than those of phase 1 models across all performance metrics except sensitivity for social work services (phase 1: 67.0%, 95% CI 63.4%-72.2%; phase 2: 72.4%, 95% CI 69.1%-75.6%; *P*=.07). Phase 2 decision models reported performance measures ranging from 59.2% to 99.3% which were statistically superior to performance measures reported by phase 1 models which ranged from 38.2% to 88.3%. For every clinical vector, phase 2 models reported significantly better area under the receiver operating characteristic curve values than those reported for phase 1 models (behavioral health: *P*=.01; social work: *P*=.03; dietitian: *P*=.001; other: *P*=.02). Furthermore, phase 2 precision scores were significantly greater than those reported in phase 1 (behavioral health: *P*<.001; social work: *P*<.001; dietitian: *P*=.02; other: *P*<.001). We also evaluated model fit using logarithmic loss (log loss), which

measures the performance of a classification model where prediction input is a probability between 0 and 1, and using lift curves [30], which compares a decision model to a random model for the given percentile of top scored predictions. Log loss values were 0.09 (behavioral health), 0.07 (social work), 0.32 (dietitian), and 0.34 (other). Lift scores for each decision model are shown in a figure in Multimedia Appendix 3.

Table 2. Characteristics of the adult, primary care patient sample whose data were used in phase 2 risk predictive modeling.

Demographic characteristics	Values
Age (years), mean (SD)	44.1 (16.6)
Gender (N=72,484), n (%)	
Male	25,297 (34.9)
Female	47,187 (65.1)
Insurance provider (N=72,484), n (%)	
Medicaid or public insurance	41,316 (57.0)
Private	31,168 (43.0)
BMI category (N=72,484), n (%)	
BMI<18.5	6379 (8.8)
18.5≤BMI<25	8698 (12.0)
25≤BMI<30	10,148 (14.0)
BMI≥30	20,875 (28.8)
Missing	26,384 (36.4)
Ethnicity (N=72,484), n (%)	
White, non-Hispanic	18,266 (25.2)
African American, non-Hispanic	34,575 (47.7)
Hispanic	15,149 (20.9)
Other	4494 (6.2)

Table 3. Comparison of clinical vector model performance for phase 1 and phase 2.

Clinical vector performance measures	Model performance, % (95% CI)		
	Phase 1	Phase 2	*P* value[a]
Behavioral health services			
Sensitivity	70.2 (68.0, 72.5)	86.3 (83.1, 88.9)	<.001
Specificity	78.5 (78.0, 78.9)	99.1 (98.5, 99.7)	<.001
F1 score	56.6 (53.6, 58.9)	90.4 (87.4, 93.4)	<.001
Precision (positive predictive value)	47.4 (44.2, 49.6)	95.0 (92.0, 98.3)	<.001
AUROC[b]	88.3 (87.4, 89.2)	98.0 (97.6, 98.5)	.01
Social work services			
Sensitivity	67.0 (63.4, 72.2)	72.4 (69.1, 75.6)	.07
Specificity	79.6 (79.1, 79.8)	99.3 (99.2, 99.6)	<.001
F1 score	48.6 (45.0, 52.5)	82.5 (79.7, 85.3)	<.001
Precision (positive predictive value)	38.2 (34.8, 41.2)	95.8 (93.8, 97.8)	<.001
AUROC	87.6 (86.1, 89.2)	93.7 (92.5, 95.0)	.03
Dietitian counseling services			
Sensitivity	60.7 (56.5, 64.7)	73.6 (70.5, 77.0)	.02
Specificity	73.2 (71.9, 74.9)	93.3 (90.8, 94.6)	<.001
F1 score	61.5 (57.3, 66.0)	76.4 (73.3, 80.4)	.001
Precision (positive predictive value)	62.2 (58.1, 67.4)	79.4 (76.4, 84.2)	.02
AUROC	82.5 (81.5, 83.6)	91.5 (90.3, 92.6)	.001
Other wraparound services			
Sensitivity	44.5 (42.7, 46.1)	59.2 (56.5, 63.8)	.002
Specificity	78.5 (77.5, 79.3)	92.9 (89.7, 96.1)	<.001
F1 score	43.2 (40.0, 45.7)	65.5 (62.9, 67.6)	.01
Precision (positive predictive value)	41.9 (37.7, 45.2)	73.4 (70.5, 77.7)	<.001
AUROC	77.2 (76.2, 78.1)	85.3 (84.4, 86.0)	.02

[a]*P* values were calculated using confidence intervals [29].

[b]AUROC: area under the receiver operating characteristic curve.

Discussion

Principal Findings

Our study expanded upon our previous efforts to demonstrate the feasibility of predicting the need for wraparound services such as behavioral health, dietitian, social work and other services using a range of readily available patient- and population-level data sets that represent an individual's well-being as well as their socioeconomic environment. Specifically, we demonstrated that inclusion of additional patient-level data sets that represented medication history, addiction and mental disorders, and patient-level social risk factors, as well as use of the extreme gradient boosting classification algorithm and advanced analytical methods for model development led to statistically superior performance measures. Furthermore, improved precision scores were made possible by additional data elements and alternate optimization techniques that maximized precision and recall scores and which greatly improved the practical application of our solution. Each decision model reported area under the receiver operating characteristic curve scores from 85% to 98%, which are superior to the global performance of prediction models on mortality [31], hospital readmissions [4], and disease development [32]; however, inclusion of additional population-level aggregate social determinants of health measures in our low-income population did not contribute significantly toward performance improvements despite the introduction of additional indicators, more granular geographic measurement units (by switching from zip code to census tract level), and vectorization methods that converted these to standardized scores to emphasize variance and create indices.

The inability of population-level social determinants of health measures to improve model performance may be because our patient population was comprised of an urban safety-net group with relatively little variability in socioeconomic, policy, and environmental conditions. Thus, it is possible that machine learning studies using larger, more diverse populations may benefit from the use of population-level data [33]. Moreover,

the lack of improvement may be related to our choice of prediction outcome. Wraparound service providers work to address the social needs and risk factors of individual patients and not population-level social determinants. Likewise, social determinants of health factors influence social risk [34], but these population conditions are not the reason for referral to a wraparound service provider. It is likely that social factors are more relevant to, and observed by, the referring provider. Nevertheless, the continued lack of meaningful contribution to our models prompts questions regarding how to best leverage aggregate social determinants of health measures for decision making. This is an important and unanswered question, as census-based aggregate measures are the most widely available and easily accessible indicators of social determinants of health available to researchers and health organizations [35]. In contrast, several patient-level social and behavioral factors measures were influential in the models. This indicates the need for more widespread use and collection of social factors in clinical settings [36]. Electronic health record organizations seeking to identify patients with social risk factors and in need of social services must integrate the collection of social risk data into their workflow [37].

Limitations

This work has limitations. Notably, the phase 2 model development approach leveraged the same urban safety-net population that was used to develop phase 1 models. Thus, though the phase 2 demonstrate superior performance, the results may not be generalizable to other commercially insured or broader populations. In addition, we only leveraged structured data that had been extracted from the Indiana Network for Patient Care or from Eskenazi Health for the machine learning process. These methods may not be utilized at other health care settings that are not part of a large, robust health information exchange. Expanding our approaches to different geographic regions would require standardization of population-level sources as well as infrastructure and interoperability measures to effectively store and exchange such data sets [38]. Also, we did not utilize any unstructured data sets for machine learning. This is a significant issue as up to 80% of health data may be collected in an unstructured format [39,40]. Despite these limitations, the considerable performance enhancements demonstrated by these models suggest significant potential to enable access to various social services; however, it must be noted that social determinants of health risk factors are often confounded with one another. Thus, mitigating a social need that arises from several social determinants of health risk factors may not result in any positive improvements to a patient [41].

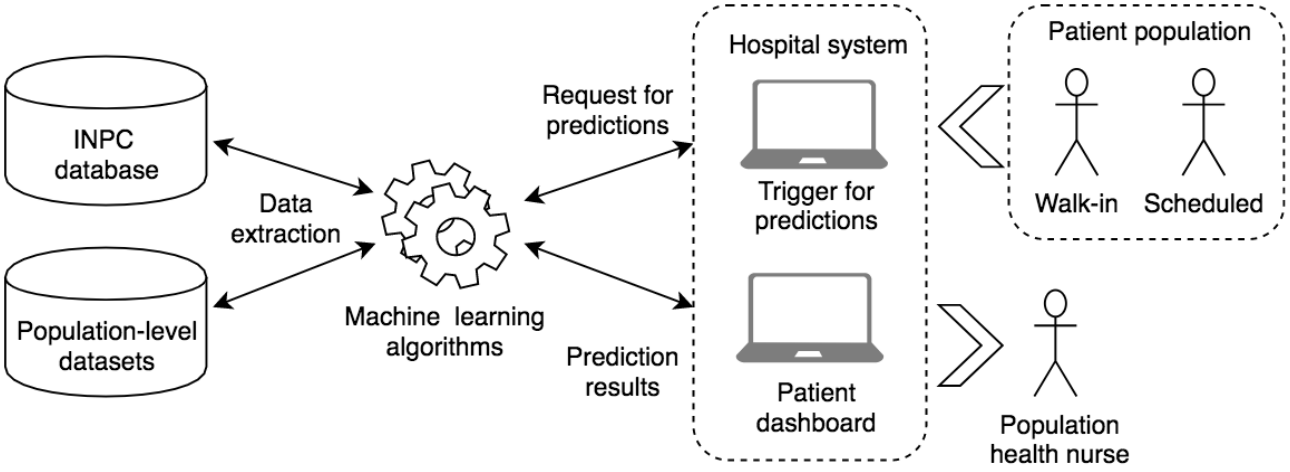
Future Work

Our next steps include expanding our models to predict additional wraparound services of interest. Furthermore, we believe that there is an acute need to improve the explainability and actionability of machine learning predictions using novel methods such as counterfactual reasoning [42]. We perceive that similar predictive models for minors and the services available to these patients would be of significant value for health care decision making. Our inability to utilize unstructured data sets for machine learning is a significant concern. Various natural language processing toolkits can leverage unstructured data sets for machine learning; however, integrating these toolkits into inproduction systems is challenging due to infrastructure and maintenance costs. Moreover, searching and indexing the massive quantities of free-text reports that are collected statewide would require additional computational effort, and may significantly increase computation time. We are currently engaged in efforts to utilize the Regenstrief Institute's nDepth tool [43] to evaluate the ability to extract actionable elements at a production setting.

Integration Into Electronic Health Record Systems

As noted, this work built upon existing risk prediction efforts. We have integrated the updated decision models into the existing platform for all scheduled and walk-in appointments. Model results are presented to end users using a customized interface within the electronic health record with metadata on which features drove the extreme gradient boosting decision-making process, and with predicted probabilities categorized as low, rising, or high risk [12] (Figure 2). This study's methodological work sets the foundation for our future evaluations of our intervention's impact on patient outcomes.

Figure 2. Integration of decision models into hospital workflow. INPC: Indiana Network for Patient Care.



Conclusions

This study developed decision models that integrate a wide range of individual and population data elements and advanced machine learning methods that are capable of predicting need for various wraparound social services; however, population-level data may not contribute to improvements in predictive performance unless they represent larger, diverse populations.

Acknowledgments

Support for this research was provided by the Robert Wood Johnson Foundation. The views expressed herein do not necessarily reflect the views of the Robert Wood Johnson Foundation. The authors also wish to thank Jennifer Williams (Regenstrief Institute), Amber Blackmon (Indiana University), the Regenstrief data core team, and Eskenazi Health of Indiana for their assistance.

Conflicts of Interest

SNK, SG, PKH, NM, and JRV are cofounders of Uppstroms LLC, a commercial entity established to disseminate the artificial intelligence models discussed in this paper.

Multimedia Appendix 1
List of risk conditions used to identify a subpopulation for predicting dietitian referrals.
[DOCX File , 14 KB - medinform_v8i7e16129_app1.docx]

Multimedia Appendix 2
Parameters that were modified as part of the hyperparameter tuning process for phase 2.
[DOCX File , 13 KB - medinform_v8i7e16129_app2.docx]

Multimedia Appendix 3
Lift scores reported by each decision model.
[PNG File , 96 KB - medinform_v8i7e16129_app3.png]

References

1. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015 Feb 26;372(9):793-795 [FREE Full text] [doi: 10.1056/NEJMp1500523] [Medline: 25635347]
2. Martin Sanchez F, Gray K, Bellazzi R, Lopez-Campos G. Exposome informatics: considerations for the design of future biomedical research information systems. J Am Med Inform Assoc 2014;21(3):386-390 [FREE Full text] [doi: 10.1136/amiajnl-2013-001772] [Medline: 24186958]
3. Charles D, Gabriel M, Searcy T. Adoption of electronic health record systems among US non-federal acute care hospitals. ONC data brief 2013;9:2008-2012 [FREE Full text]
4. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011 Oct 19;306(15):1688-1698 [FREE Full text] [doi: 10.1001/jama.2011.1515] [Medline: 22009101]
5. Vuik SI, Mayer EK, Darzi A. Patient segmentation analysis offers significant benefits for integrated care and support. Health Aff (Millwood) 2016 May 01;35(5):769-775. [doi: 10.1377/hlthaff.2015.1311] [Medline: 27140981]
6. Vest JR, Harris LE, Haut DP, Halverson PK, Menachemi N. Indianapolis provider's use of wraparound services associated with reduced hospitalizations and emergency department visits. Health Aff (Millwood) 2018 Oct;37(10):1555-1561. [doi: 10.1377/hlthaff.2018.0075] [Medline: 30273041]
7. Lewis JH, Whelihan K, Navarro I, Boyle KR, SDH Card Study Implementation Team. Community health center provider ability to identify, treat and account for the social determinants of health: a card study. BMC Fam Pract 2016 Aug 27;17:121 [FREE Full text] [doi: 10.1186/s12875-016-0526-8] [Medline: 27567892]
8. Fenton M. Health care's blind side: the overlooked connection between social needs and good health. Princeton: Robert Wood Johnston Foundation 2011.
9. Fitzpatrick T, Rosella L, Calzavara A, Petch J, Pinto A, Manson H, et al. Looking beyond income and education: socioeconomic status gradients among future high-cost users of health care. Am J Prev Med 2015 Aug;49(2):161-171 [FREE Full text] [doi: 10.1016/j.amepre.2015.02.018] [Medline: 25960393]
10. Kaufman A. Theory vs practice: should primary care practice take on social determinants of health now? yes. Ann Fam Med 2016 Mar;14(2):100-101 [FREE Full text] [doi: 10.1370/afm.1915] [Medline: 26951582]
11. Kasthurirathne S, Vest J, Menachemi N, Halverson P, Grannis S. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. J Am Med Inform Assoc 2018 Jan 01;25(1):47-53. [doi: 10.1093/jamia/ocx130] [Medline: 29177457]

12. Vest JR, Menachemi N, Grannis SJ, Ferrell JL, Kasthurirathne SN, Zhang Y, et al. Impact of risk stratification on referrals and uptake of wraparound services that address social determinants: a stepped wedged trial. Am J Prev Med 2019 Apr;56(4):e125-e133. [doi: 10.1016/j.amepre.2018.11.009] [Medline: 30772150]
13. ICD-10 adds more detail on the social determinants of health. LaBrec P. 2016. URL: https://www.3mhisinsideangle.com/blog-post/icd-10-adds-more-detail-on-the-social-determinants-of-health/ [accessed 2019-02-24] [WebCite Cache ID 76Qh0qs84]
14. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]
15. Choi E, Bahadori M, Schuetz A, Stewart W, Sun J. Doctor ai: predicting clinical events via recurrent neural networks. arXiv preprint arXiv 2015:151105942.
16. Duke JD, Morea J, Mamlin B, Martin DK, Simonaitis L, Takesue BY, et al. Regenstrief Institute's Medical Gopher: a next-generation homegrown electronic medical record system. Int J Med Inform 2014 Mar;83(3):170-179. [doi: 10.1016/j.ijmedinf.2013.11.004] [Medline: 24373714]
17. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, INPC Management Committee. The Indiana network for patient care: a working local health information infrastructure. Health Aff (Millwood) 2005;24(5):1214-1220. [doi: 10.1377/hlthaff.24.5.1214] [Medline: 16162565]
18. Goodman RA, Posner SF, Huang ES, Parekh AK, Koh HK. Defining and measuring chronic conditions: imperatives for research, policy, program, and practice. Prev Chronic Dis 2013 Apr 25;10:E66 [FREE Full text] [doi: 10.5888/pcd10.120239] [Medline: 23618546]
19. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc 2011 Jul 01;18(4):441-448. [doi: 10.1136/amiajnl-2011-000116] [Medline: 21515544]
20. Beyond health care: the role of social determinants in promoting health and health equity. Artiga S, Hinton E. 2018. URL: http://files.kff.org/attachment/issue-brief-beyond-health-care [accessed 2020-06-03]
21. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 2010 Jun;31(8):651-666. [doi: 10.1016/j.patrec.2009.09.011]
22. Kodinariya T, Makwana P. Review on determining number of Cluster in K-Means Clustering. International Journal 2013;1(6):90-95.
23. Tang J, Alelyani S, Liu H. Feature selection for classification: a review. In: Aggarwal CC, editor. Data Classification: Algorithms and Applications. New York: Chapman and Hall/CRC; Jul 25, 2014:A.
24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. (Oct) 2011;12:2825-2830 [FREE Full text]
25. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016 Aug 1 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2016; San Francisco.
26. Ye J, Chow J, Chen J, Zheng Z, editors. Stochastic gradient boosted distributed decision trees. 2019 Dec 1 Presented at: Proceedings of the 18th ACM conference on Information and Knowledge management; : ACM; 2009; Hong Kong. [doi: 10.1145/1645953.1646301]
27. Nielsen D. Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition? In: NTNU Open. Trondheim: NTNU; 2016.
28. Calonico S, Cattaneo MD, Titiunik R. Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica 2014 Dec 23;82(6):2295-2326. [doi: 10.3982/ecta11757]
29. Altman DG, Bland JM. How to obtain the P value from a confidence interval. BMJ 2011;343:d2304. [doi: 10.1136/bmj.d2304] [Medline: 22803193]
30. Vuk M, Curk T. ROC curve, lift chart and calibration plot. Metodoloski zvezki 2006;3(1):89.
31. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, et al. Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. Circ Heart Fail 2013 Sep 01;6(5):881-889. [doi: 10.1161/CIRCHEARTFAILURE.112.000043] [Medline: 23888045]
32. Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk models to predict hypertension: a systematic review. PLoS One 2013;8(7):e67370 [FREE Full text] [doi: 10.1371/journal.pone.0067370] [Medline: 23861760]
33. Dalton JE, Perzynski AT, Zidar DA, Rothberg MB, Coulton CJ, Milinovich AT, et al. Accuracy of cardiovascular risk prediction varies by neighborhood socioeconomic position: a retrospective cohort study. Ann Intern Med 2017 Oct 03;167(7):456-464 [FREE Full text] [doi: 10.7326/M16-2543] [Medline: 28847012]
34. Alderwick H, Gottlieb LM. Meanings and misunderstandings: a social determinants of health lexicon for health care systems. Milbank Q 2019 Jun;97(2):407-419 [FREE Full text] [doi: 10.1111/1468-0009.12390] [Medline: 31069864]
35. Golembiewski E, Allen KS, Blackmon AM, Hinrichs RJ, Vest JR. Combining nonclinical determinants of health and clinical data for research and evaluation: rapid review. JMIR Public Health Surveill 2019 Oct 07;5(4):e12846 [FREE Full text] [doi: 10.2196/12846] [Medline: 31593550]
36. Institute of Medicine. National Academies Press (US). Capturing social and behavioral domains and measures in electronic health records: phase 2: National Academies Press; 2014:0309312434.

37. Gold R, Bunce A, Cowburn S, Dambrun K, Dearing M, Middendorf M, et al. Adoption of social determinants of health EHR tools by community health centers. Ann Fam Med 2018 Sep;16(5):399-407 [FREE Full text] [doi: 10.1370/afm.2275] [Medline: 30201636]
38. Kasthurirathne S, Cormer K, Devadasan N, Biondich P. editors. 2019 Mar 25 Presented at: Development of a FHIR Based Application Programming Interface for Aggregate-Level Social Determinants of Health. : AMIA Informatics summit Conference Proceedings; 2019; San Francisco. [doi: 10.4135/9781529705119]
39. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change 2018 Jan;126:3-13. [doi: 10.1016/j.techfore.2015.12.019]
40. Hubbard WN, Westgate C, Shapiro LM, Donaldson RM. Acquired abnormalities of the tricuspid valve--an ultrasonographic study. Int J Cardiol 1987 Mar;14(3):311-318. [doi: 10.1016/0167-5273(87)90201-4] [Medline: 3549579]
41. Castrucci B, Auerbach J. Meeting individual social needs falls short of addressing social determinants of health. Health Affairs Blog. URL: https://www.healthaffairs.org/do/10.1377/hblog20190115.234942/full/?utm_campaign=HASU&utm_medium=email&utm_content=Health+Affairs+In+2018%3A+Editor+s+Picks%3B+The+2020+Proposed+Payment+Notice%3B+Persistently+High-Cost+Medicare+Patients&utm_source=Newsletter& [accessed 2019-01-01]
42. Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv 2018:180104016. [doi: 10.1145/3159652.3176182]
43. Regenstrief Institute Inc.. nDepth. regenstrief.org. 2019. URL: https://www.regenstrief.org/implementation/ndepth/ [accessed 2019-01-01]

Abbreviations

AUROC: area under the receiver operating characteristic curve
ICD-10: International Statistical Classification of Diseases, 10th Revision

Edited by G Eysenbach; submitted 08.09.19; peer-reviewed by K Davison, I ten Klooster, A Sheon, K Mandl; comments to author 12.12.19; revised version received 03.02.20; accepted 09.04.20; published 09.07.20.

Please cite as:
Kasthurirathne SN, Grannis S, Halverson PK, Morea J, Menachemi N, Vest JR
Precision Health–Enabled Machine Learning to Identify Need for Wraparound Social Services Using Patient- and Population-Level Data Sets: Algorithm Development and Validation
JMIR Med Inform 2020;8(7):e16129
URL: https://medinform.jmir.org/2020/7/e16129
doi:10.2196/16129
PMID:32479414

Original Paper

Identifying the Medical Lethality of Suicide Attempts Using Network Analysis and Deep Learning: Nationwide Study

Bora Kim[1*], MD, MAS; Younghoon Kim[2*], PhD; C Hyung Keun Park[3,4], MD, PhD; Sang Jin Rhee[3,4], MD; Young Shin Kim[1], MS, MPH, MD, PhD; Bennett L Leventhal[1], MD; Yong Min Ahn[3,4*], MD, PhD; Hyojung Paik[2*], PhD

[1]Department of Psychiatry, University of California, San Francisco, San Francisco, CA, United States

[2]Center for Supercomputing Applications, Division of Supercomputing, Korea Institute of Science and Technology Information (KISTI), Daejeon, Republic of Korea

[3]Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Republic of Korea

[4]Department of Psychiatry and Behavioral Science, Seoul National University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hyojung Paik, PhD
Center for Supercomputing Applications
Division of Supercomputing
Korea Institute of Science and Technology Information (KISTI)
245 Daehak-ro
Yuseong-gu
Daejeon, 305-806
Republic of Korea
Phone: 1 82 42 869 1004
Email: hyojungpaik@kisti.re.kr

Abstract

Background: Suicide is one of the leading causes of death among young and middle-aged people. However, little is understood about the behaviors leading up to actual suicide attempts and whether these behaviors are specific to the nature of suicide attempts.

Objective: The goal of this study was to examine the clusters of behaviors antecedent to suicide attempts to determine if they could be used to assess the potential lethality of the attempt. To accomplish this goal, we developed a deep learning model using the relationships among behaviors antecedent to suicide attempts and the attempts themselves.

Methods: This study used data from the Korea National Suicide Survey. We identified 1112 individuals who attempted suicide and completed a psychiatric evaluation in the emergency room. The 15-item Beck Suicide Intent Scale (SIS) was used for assessing antecedent behaviors, and the medical outcomes of the suicide attempts were measured by assessing lethality with the Columbia Suicide Severity Rating Scale (C-SSRS; lethal suicide attempt >3 and nonlethal attempt ≤3).

Results: Using scores from the SIS, individuals who had lethal and nonlethal attempts comprised two different network nodes with the edges representing the relationships among nodes. Among the antecedent behaviors, the conception of a method's lethality predicted suicidal behaviors with severe medical outcomes. The vectorized relationship values among the elements of antecedent behaviors in our deep learning model (E-GONet) increased performances, such as F1 and area under the precision-recall gain curve (AUPRG), for identifying lethal attempts (up to 3% for F1 and 32% for AUPRG), as compared with other models (mean F1: 0.81 for E-GONet, 0.78 for linear regression, and 0.80 for random forest; mean AUPRG: 0.73 for E-GONet, 0.41 for linear regression, and 0.69 for random forest).

Conclusions: The relationships among behaviors antecedent to suicide attempts can be used to understand the suicidal intent of individuals and help identify the lethality of potential suicide attempts. Such a model may be useful in prioritizing cases for preventive intervention.

(JMIR Med Inform 2020;8(7):e14500) doi:10.2196/14500

KEYWORDS

suicide; deep learning; network; antecedent behaviors

Introduction

Suicide is an important public health epidemic globally. The suicide incidence in the United States has increased in recent years from 10.9/100,000 in 2006 to 13.3/100,000 in 2015 [1], and nearly 45,000 Americans killed themselves in 2016 [2]. The suicide rate in South Korea is the highest among developed countries, and mortality attributable to suicide exceeds that attributable to common diseases, including diabetes, pneumonia, and liver disease [3]. Suicide is a preventable health problem, but effective prevention strategies are lacking because it is a complex issue, and thus, it is difficult for researchers to develop a cause and prediction model [4].

The management of suicide attempts is an urgent clinical problem, and preventing further attempts is particularly important. The risk of suicide has many components, and of these, a previous suicide attempt is among the most important [5,6]. Understanding the nature of suicide attempts and possible associations with subsequent death by suicide may facilitate the design of interventions targeted at specific risk characteristics for particular individuals, thereby increasing clinical effectiveness and reducing morbidity and mortality in this high-risk population.

Suicide attempts are highly heterogeneous and range from a "cry for help" to a nearly lethal attempt with self-mutilation and actual suicide [7]. In the present study, among the outcomes of suicide attempts, we consider the possible medical lethality of attempts as medical consequences, as well as the severity of the physical harm to individuals. Medical lethality as an outcome can be considered the degree of danger to life resulting from a suicide attempt [8]. In addition, most people who attempt suicide will communicate their intent in various forms before they actually attempt suicide [9]. However, it is unclear whether understanding the specific relationships with the behaviors leading up to actual suicide attempts can help to provide guidance for reducing suicide attempts. The relationship between the lethality of a preceding suicide attempt and medical lethality following a subsequent suicide attempt is unknown [10]. Thus, predictive models and explorations of the thought structures of individuals who attempt suicide are still lacking.

We hypothesized that among individuals who attempt suicide, the relationships among their antecedent suicidal thoughts, behaviors, and communications will exhibit specific patterns, thereby allowing us to predict their future risk and lethality. A network model can be employed to conceptualize the complex dynamic systems comprising each interacting symptom [11-13]. Owing to this advantage of network analysis, previous studies have explored the nested interactions among the features of psychopathology [14] or the symptoms of major depressive disorder [15]. The results obtained by network-based analysis can successfully depict multiple nodes as variables, and multiple edges represent the mutual interactions between each pair of variables (ie, nodes). In this study, we employed network analysis to build a model where the nodes represent unique aspects of the expressed suicidal intent and the edges depict the correlations among these nodes.

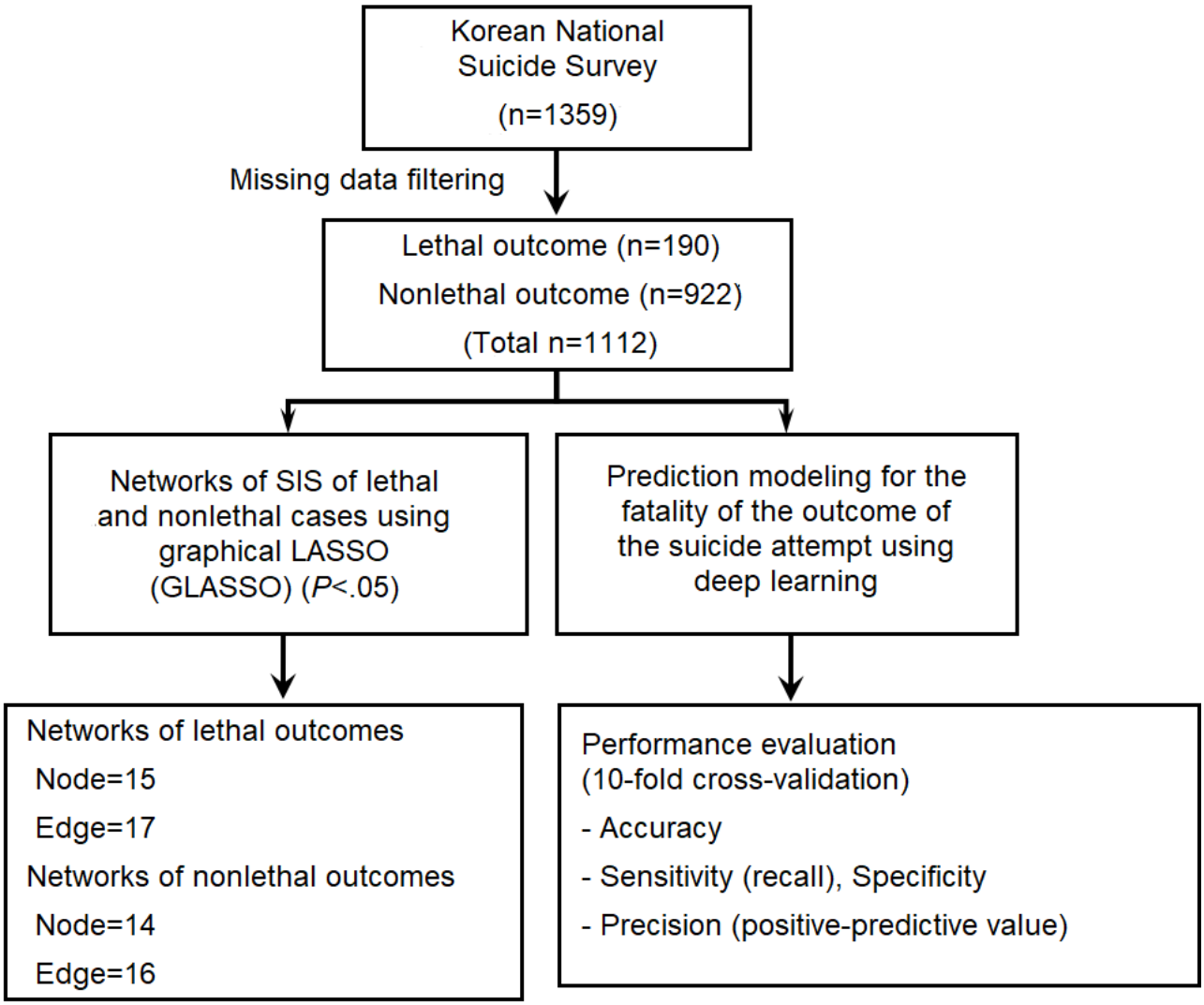
After determining the relationship values among the antecedent behaviors, we applied deep learning to identify the medical outcomes of subsequent suicide attempts. Deep learning is an emerging machine learning technique for predictive modeling in various applications, which is based on data observations but without domain-specific knowledge. The application of deep learning in the psychiatric field to develop Woebot, a text-based chatbot, has facilitated depression care [16]. However, the success of machine learning-based approaches has been limited in the identification of the central elements of suicidal intents and in prediction modeling based on the information collected by health care providers to meaningfully enhance clinical care. In this study, we employed a network-based method to explore the connections between communicative behaviors prior to suicide attempts with lethal or less lethal outcomes by using data that are routinely collected by physicians. Moreover, to train the complex connections between the antecedent behaviors, our deep learning model utilized the novel relationship values among suicidal intent elements.

Methods

Study Sample

We analyzed data obtained from the Korea National Suicide Survey [17], which was a nationwide multicenter study of subjects from two cohorts comprising individuals who attempted suicide and were recruited by retrospective chart review and those who attempted suicide and completed psychiatric evaluations by on-call psychiatric residents. The subjects from the second dataset were used in this study. All individuals who attempted suicide visited the emergency room (ER), and they were evaluated in semistandardized interviews at 17 medical centers across South Korea from May 1, 2013, to November 7, 2013. Deaths in the ER were excluded from the data. Among 1359 individuals who attempted suicide, 1112 were included in the final analysis after excluding missing data from the Columbia Suicide Severity Rating Scale (C-SSRS) and Beck Suicide Intent Scale (SIS) (Figure 1).

Figure 1. Study overview. C-SSRS: Columbia Suicide Severity Rating Scale; Edge: Association between a pair of nodes based on the weighted correlations according to graphical lasso; Node: Nodes for the measured elements of the SIS and C-SSRS fatality assessment; SIS: Beck Suicide Intent Scale.



Outcome: Medical Lethality of Suicide Attempts

The outcome of this study involving the medical lethality of suicide attempts was assessed by a clinician and classified based on the "actual lethality or medical damage" using the C-SSRS. The validated Korean version of the C-SSRS was used [18], and lethality was rated as follows: 1, no physical damage or very minor physical damage (eg, surface scratches); 2, minor physical damage (eg, lethargic speech and mild bleeding); 3, moderate physical damage (eg, conscious but sleepy); 4, moderately severe physical damage (eg, comatose with reflexes); 5, severe physical damage (eg, comatose without reflexes); and 6, death [19]. We used a lethality scale with the following two categories: score 3, a less lethal outcome of a suicide attempt and score >3, a lethal outcome of a suicide attempt.

Suicidal Intents: Suicidal Intent Thoughts, Behaviors, and Communications

The SIS was used to identify the elements of suicidal intent thoughts, behaviors, and communications [20]. The scale contains 15 questions (ie, SIS 1-15), and all of the items are scored on a scale from 0 to 2 for severity, where the total sum of the scores ranges from 0 to 30. In this study, we calibrated the SIS scale from 1 to 3 to calculate the relationships among the features. The SIS comprises the following two parts: objective circumstances of the attempt and the subject's self-reported intentions and expectations regarding the attempt. The objective factors (SIS 1-8) are as follows: SIS 1, isolation; SIS 2, timing of intervention feasibility; SIS 3, active or passive precautions against discovery or intervention; SIS 4, acting to get help during or after the suicide attempt; SIS 5, final acts in anticipation of death; SIS 6, active preparation for the suicide attempt; SIS 7, suicide note; and SIS 8, overt communication of intent before the suicide attempt. The subjective factors (SIS 9-15) are as follows: SIS 9, alleged purpose of the suicide attempt; SIS 10, expectations of fatality; SIS 11, conception of a method's lethality; SIS 12, seriousness of the suicide attempt; SIS 13, attitude toward living or dying; SIS 14, conception of medical rescuability; and SIS 15, degree of premeditation.

Utilization and Reprocessing of Confounders in the ER

We also used clinical data reported from the ER as confounding variables in prediction modeling. In total, 14 confounders were

considered, including sex, age, marital status, religion, monthly income, living status, educational level, urbanicity, ER visit date, ER visit on a weekend, ER visit time, admission route, admission transportation, and discharge date. All of the confounders were collected as numerical values, with coded indices or quantitative values as follows: sex (1=male, 2=female), marital status (1=single, 2=married, 3=living together, 4=separated, 5=divorced, 6=widowed), religion (1=Christian, 2=Buddhist, 3=Catholic, 4=Atheist, 5=other), living status (1=living with family, 2=living with somebody, 3=group facilities, 4=living alone), educational level (1=none, 2=elementary school, 3=middle school, 4=high school, 5=undergraduate or higher), ER visit on a weekend (1=yes, 2=no), admission route, admission transportation, and monthly income (self-reported in Korean currency). To represent the date records numerically (ie, year-month-day, ER visit date, and discharge date), we transformed the original date into a decimal year value (eg, 2013-6-30=2013.492). The detailed equation for the numerical transformation of the dates is presented in the supplementary source code [21].

Relationships Among Suicidal Intent Items

For each pair of 15 suicidal intent items for each individual, we generated three relationship signatures comprising the interaction terms (*I*), harmonized average (*H*), and geometric angle differences using the tangent function (*T*). The definitions of the three relationships between the *ith* and *jth* element of SIS in the *pth* individual ($R(I, H, T)$) are represented by the following equations:

where S_i^p indicates the *ith* SIS element in the *pth* individual, and $I_{i,j}^p$ determines the level of interaction between the *ith* and *jth* SIS elements. To represent the overall intensity in a sensitive manner, we utilized the harmonic mean of a pair of elements ($H_{i,j}^p$). According to the differences in the sequential combination of a pair of elements, such as $[S_i^p=2 > S_j^p=3]$ and $[S_i^p=3< S_j^p=2]$, $T_{i,j}^p$ presents a single scalar value for the paired elements.

Data Analysis

The chi-square test and Student *t* test were performed to compare variables for suicide attempts with lethal and nonlethal medical outcomes.

Missing Data Imputation

The k-nearest neighbors algorithm in the R package bnstruct [22] was utilized to impute any missing values (the proportion of missing values in our data was 2.7%).

Network Analysis

Network model analysis was performed to build a relational model of lethal and nonlethal suicide attempts. Using the graphical lasso (GLASSO) method, we investigated the weighted correlations between the assessed SIS elements according to attempt fatalities [23]. The network comprised nodes representing the suicidal intent elements, and the edges depicted the relationships among nodes as the medical outcomes of those who had lethal and nonlethal suicide attempts (Figure 1). The statistical significance levels of the GLASSO results were determined using the random 10,000-permutation method ($P<.05$). R [22] and Cytoscape [24] were employed for data analysis and visualization of the results, respectively. The source code was deposited in the GitHub database [21].

Machine Learning

Machine learning techniques comprising random forest and linear regression were used. To evaluate the contributions of the relationship scores, we compared the predictive performances of models with or without the relationship features. We utilized TensorFlow [25] to develop our deep learning model called E-GONet. In addition, the feature importance map for the input data of convolutional neural network (CNN) models was generated by DeepExplain with the “Gradient*Input” method [26]. To obtain the feature map, the feature importance scores of all nonlethal cases and all lethal cases were averaged in each fold of 10-fold cross validation sets, and then, the average scores in each fold were averaged into final feature importance scores for nonlethal and lethal cases, respectively.

Results

Characteristics of Lethal and Nonlethal Outcomes of Suicide Attempts

Among 1112 individuals who attempted suicide, 190 (17.1%) had suicide attempts categorized as lethal medical outcomes (Table 1). According to the C-SSRS–based fatality of attempt outcomes, we classified the individuals as those who had lethal attempts (n=190, C-SSRS severity 3) and those who had nonlethal attempts (n=922). More male individuals had lethal suicide attempts than nonlethal suicide attempts (107/190, 56.3% vs 357/922, 38.7%). The mean age of those who had lethal suicide attempts was higher than that of those who had nonlethal suicide attempts (47.3 years vs 42.3 years).

The mean total SIS score was higher for those who had lethal suicide attempts than those who had nonlethal suicide attempts (30.23, SD 6.19 vs 25.06, SD 5.61). The total SIS score is the sum of the 15 graded elements of the SIS, such as the scale of isolation from 1 to 3 (complete isolation).

Table 1. Demographic characteristics.

Features	Total[a] (n=1112)	Lethal attempts[a,b] (n=190)	Nonlethal attempts[a,b] (n=922)	*P* value
Sex				<.001[c]
Female	647 (58.18)	83 (43.68)	564 (61.17)	
Male	464 (41.73)	107 (56.32)	357 (38.72)	
Unknown	1 (0.09)	0 (0.0)	1 (0.11)	
Mean age, years	43.17 (18.2)	47.31 (18.41)	42.31 (18.06)	<.001[d]
History of suicide attempts				.88[c]
No attempt or unknown	763 (68.62)	129 (67.89)	634 (68.76)	
Previous history	349 (31.38)	61 (32.11)	288 (31.24)	
C-SSRS[e] fatality rating for the outcome of an attempt				
1: None or minor physical damage	177 (15.92)	N/A[f]	177	
2: Minor physical damage	387 (34.80)	N/A	387	
3: Moderate physical damage	358 (32.19)	N/A	358	
4: Severe physical damage	152 (13.67)	152 (80.00)	N/A	
5: Very severe physical damage	33 (2.97)	33 (17.37)	N/A	
6: Death	5 (0.45)	5 (2.63)	N/A	
Mean of the SIS[g] sum	26.74 (5.91)	30.23 (6.19)	25.06 (5.61)	<.001[d]

[a]Data are presented as n (%), n, or mean (SD).

[b]Fatality scale of the Columbia Suicide Severity Rating Scale (3 for a lethal suicide attempt and <3 for a nonlethal suicide attempt).

[c]Chi-square test between those who had lethal attempts and those who had nonlethal attempts.

[d]*t* test between those who had lethal attempts and those who had nonlethal attempts.

[e]C-SSRS: Columbia Suicide Severity Rating Scale.

[f]N/A: not applicable.

[g]SIS: Beck Suicide Intent Scale.

Network Model Based on Suicidal Intent Elements

Using GLASSO, we determined the weighted correlations among the assessed SIS elements according to suicide attempt fatalities [23]. We constructed two networks comprising nodes representing the SIS elements (eg, degree of isolation) and edges depicting the relationships among nodes, which indicated the distinct relationships between the elements of suicide in those who had lethal and those who had nonlethal suicide attempts (Figure 1). The statistical significance of the correlations assessed among the SIS elements using GLASSO were determined based on random distributions of the SIS elements (P<.05 for random distributions). Finally, we represented the distinct relationships between the suicidal intents of those who attempted suicide (lethal and nonlethal cases).

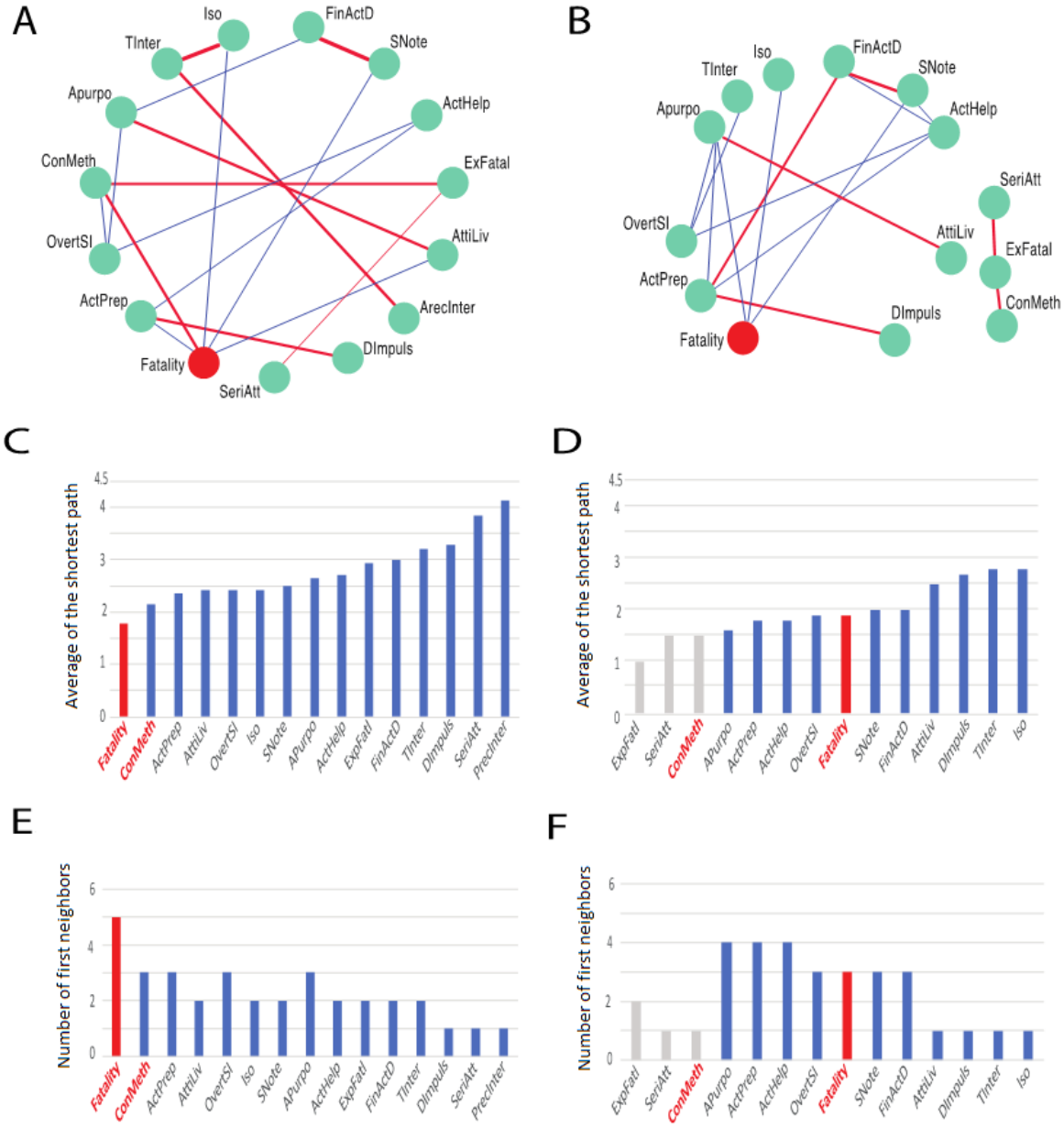
Fifteen nodes for suicidal intents were linked via 17 edges for lethal suicide attempts (n=190) (Figure 2A). Among the 922 individuals who had nonlethal suicide attempts, there were 16 relationships (ie, edges) among 14 suicidal intent elements (ie, nodes) in the nonlethal suicide attempts (Figure 2B). The edges between nodes represent the positive or negative relationships between nodes based on the GLASSO results (P<.05 for edges based on a random distribution). Among individuals who had lethal attempts, the fatal outcomes of suicide attempts were more tightly linked (ie, associated) with the suicidal intents compared with those who had nonlethal attempts, and the nodes for the SIS elements were loosely connected or separated in those who had nonlethal attempts (Figure 2A and B). Thus, the close connectedness of the suicidal intent elements, including the concept of lethality of the method, was stronger among those who had lethal attempts.

The topological properties of the network, such as the central node that had the largest number of relationships with other nodes and the average of the shortest paths (ie, degree of centrality), were employed to determine the central suicidal intent elements for the lethal and nonlethal attempts (Figure 2C-F). Among the lethal suicide attempts, the *fatality of suicide attempt* (“Fatality” node) and the *conception of a suicide method’s lethality* (“ConMeth” node) were strongly associated with other suicidal intent elements (Figure 2E). In Figure 2C, the y-axis denotes the average shortest path, which was a bottleneck and a central node, and the “ConMeth” node was ranked highly among those who had lethal attempts (Figure 2C). However, the *alleged purpose of suicide attempt* (“APurpo” node), including “to manipulate the environment,” was a crucial intent element in the minds of those who had nonlethal attempts (Figure 2D and F). Moreover, among those who had nonlethal attempts, suicide method-related features (eg, the conception

of a method's lethality and the expectation of fatality), which are closely linked to lethality among those who had lethal attempts, were completely disconnected from the other suicidal intent nodes (Figure 2D and F).

Thus, we elucidated the relationships between the suicidal intent elements in those who had lethal and those who had nonlethal suicide attempts. The *conception of a method's lethality* ("ConMeth" node) was a central suicidal intent element, which was clearly related to lethal suicide attempts, and it was connected with the initiation of attempts, such as the nodes for *expectation of fatality* ("ExFatal" node) and *seriousness of attempt* ("SeriAtt" node).

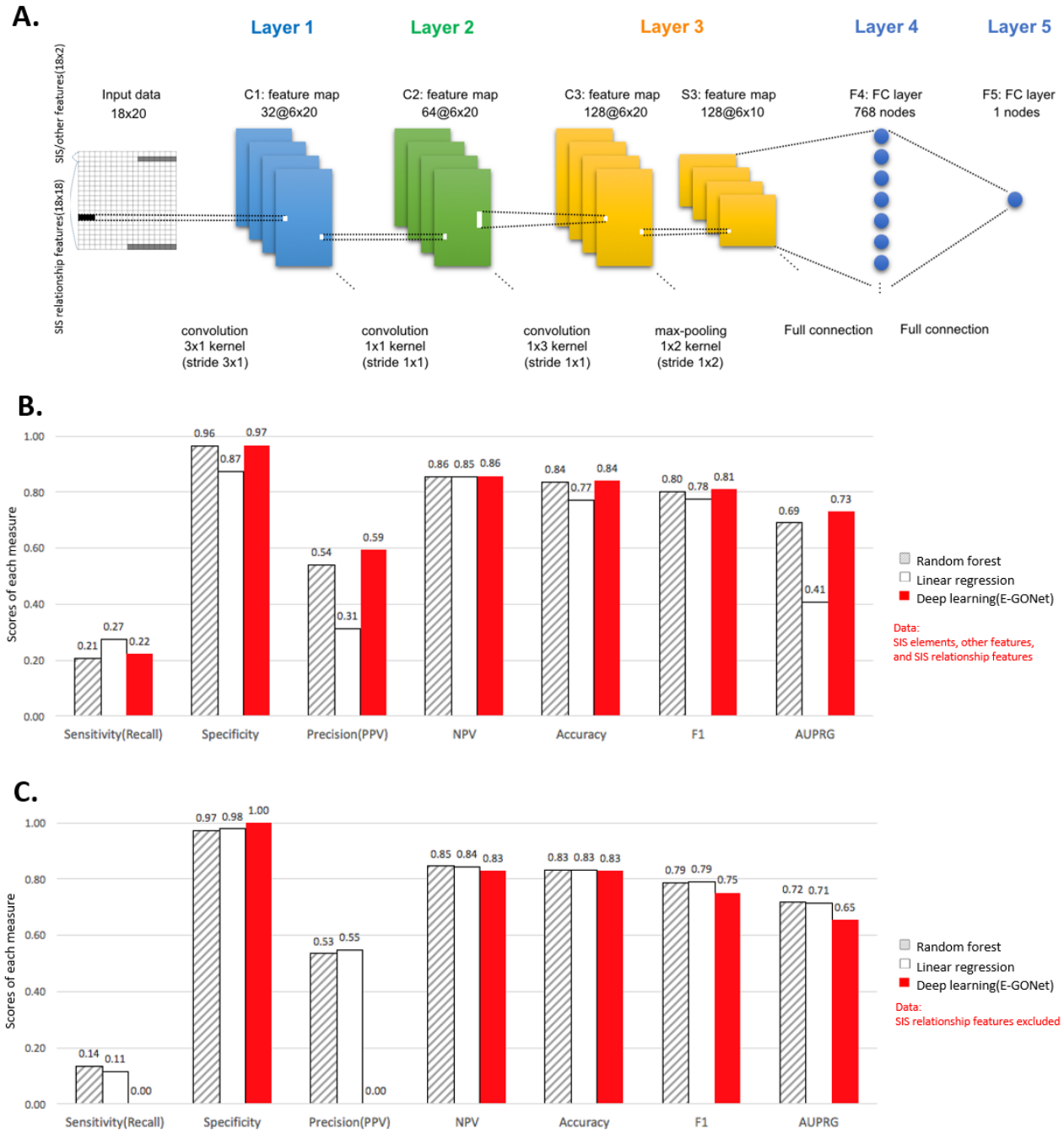
Figure 2. Network structures obtained for lethal and nonlethal outcomes. (A, B) Networks obtained for lethal (A; n=190, C-SSRS fatality ≥3) and nonlethal (B; n=922, C-SSRS fatality <3) cases. Each node represents the SIS element (green circles) and assessed fatality of the attempt using the C-SSRS score (red circles). The linked edges indicate strong relationships between nodes based on the weighted correlations obtained by graphical lasso (GLASSO) (P<.05). The red edges represent positive relationships, and the blue edges represent negative relationships. The cyan circular nodes indicate SIS elements, and the red circular nodes indicate the C-SSRS fatality scores for suicide attempts. (C-F) Bar charts showing the topological properties of the networks obtained for lethal (C, E) and nonlethal cases (D, F). The SIS elements were as follows: isolation (Iso), time intervention feasibility (Tinter), active or passive precautions against discovery intervention (ArecInter), acting to get help (ActHelp), final acts in anticipation of death (FinActD), active preparation for attempt (ActPrep), suicide note (SNote), overt communication of suicidal intent (OvertSI), alleged purpose of attempt (Apurpo), expectation of fatality (ExFatal), conception of method lethality (ConMeth), seriousness of attempt (SeriAtt), attitude toward living or dying (AttiLiv), conception of medical rescuability (ConResc), and degree of premeditation impulsiveness (Dimpuls). C-SSRS: Columbia Suicide Severity Rating Scale; SIS: Beck Suicide Intent Scale.



Predictive Model for Medical Lethality of Suicide Attempts Based on Deep Learning: E-GONet

Owing to the structural differences in the networks of antecedent behaviors according to the lethal or nonlethal outcomes of the suicide attempts, we generated three relationship signatures for each pair of SIS elements comprising the interaction terms (*I*), harmonized average (*H*), and geometric angle differences using the tangent function (*T*). In addition to the 15 SIS elements and 14 types of clinically reported data collected by the ER, including age, admission date, and living status, three relationship signatures were prepared for all possible SIS combinations for each individual. We represented the pairs of SIS elements as specific numeric values, and 315 relationship features were obtained among the 105 combinations of SIS elements for an individual. We built E-GONet based on a CNN, which is a subclass of deep learning. The overall structure of E-GONet comprises input and output layers, as well as multiple hidden layers (convolutional layers, pooling layers, and fully connected layers). The schematic structure of each layer is shown in Figure 3A. Multimedia Appendix 1 and Multimedia Appendix 2 describe the detailed structures of the E-GONet model.

Figure 3. Construction of E-GONet and performance evaluation. (A) Structure of E-GONet based on a convolutional neural network model. The input data format was 18 × 20 (row × column), which comprised SIS or other features and SIS relationship features. The SIS or other features comprised 29 features (15 SIS features and 14 observations collected by emergency rooms) and seven blanks with all zero values (gray). The SIS relationship features comprised 315 features (105 relationships × three types) and nine blanks. E-GONet has three convolutional layers and two fully connected layers. TensorFlow 1.8.0 was used for the implementation. (B) Mean performance based on 10-fold cross-validation using all of the implemented features in A. The red bar shows the average performance of E-GONet with the 10-fold cross-validation set. (C) Mean performance using the data set without SIS relationship features. AUPRG: area under the precision-recall gain curve; F1: weighted-F1 score; NPV: negative-predictive value; PPV: positive-predictive value; SIS: Beck Suicide Intent Scale.



Evaluation of E-GONet for Identifying the Medical Lethality of Suicide

Figure 3B shows the results obtained from the performance evaluation. We evaluated the predictive performance of the E-GONet model by 10-fold cross-validation. We used the same dataset for the performance comparison with E-GONet and to establish two prediction models with linear regression and random forest (ie, an aggregated decision tree model) [27] methods.

E-GONet performed better than the linear regression and random forest methods (E-GONet increased the F1 score up to 3.4%, and the mean increase in the F1 score was 2.1%; E-GONet also increased the area under the precision-recall gain curve [AUPRG] up to 32.1%, and the mean increase in the AUPRG was 18.1%) [28]. Besides, the positive-predictive value (PPV;

precision) comprising the rate of correctly identifying lethal attempts was highest with the E-GONet predictions (0.59). As generally noted in the clinical field, our dataset was relatively imbalanced (lethal 190, nonlethal 922). In analysis involving imbalanced data, sensitivity, PPV, F1, and AUPRG have been used for performance evaluation instead of specificity, negative-predictive value, and accuracy.

The analysis of the contribution of learning features (Multimedia Appendix 3 and Multimedia Appendix 4) showed that most of the contributions of confounding variables (such as level of education) were negligible for the predictive performance of E-GONet. However, the relationship signatures between the SIS element pairs contributed greatly to the superior performance of the E-GONet model (the values of $R(I, H, T)$). As depicted in Multimedia Appendix 3 and Multimedia Appendix 4, the saliency heatmaps of our model highlighted the contribution of SIS relationship features. The first two rows of the feature importance matrixes were the confounders (age, sex, income levels, etc) and SIS elements (SIS 1-15). Out of those features, only age and income level contributed to E-GONet training. On the other hand, as presented in the figures, the developed relationship features of SIS elements were more important for CNN model training. Interestingly, out of all relationships among SIS elements, the relationship with SIS element 11 (conception of a suicide method's lethality) was the biggest contributor to predictions. This is highly consistent with the network modeling of SIS elements in Figure 2.

Moreover, Figure 3C shows the evaluations of the performance of the models established without the SIS-based relationship features. The predictive models based on linear regression and random forest exhibited similar or lower performance after introducing the SIS-based relationship features, because these classical methods could not patternize the relationship features of suicide elements. However, E-GONet trained and improved performance via the vectorized relationships in high-dimensional spaces. As a result, the relationships between SIS elements increased the performance of the E-GONet model by 60% in precision, 6% in F1, and 8% in AUPRG (precision [without relationship/with relationship]=0.0/0.59, F1=0.75/0.81, and AUPRG=0.65/0.73). The full spectrum of the AUPRG displayed the training and fitting process of our deep learning approach in a very detailed manner (Multimedia Appendix 5). In Multimedia Appendix 6, the standard deviations of AUPRGs are presented via 100 trials of 10-fold cross-validation settings.

Therefore, identifying the lethality of attempt outcomes is feasible with deep learning through the major contributions of the relationships among SIS elements (ie, mutual interactions of antecedent behaviors). To allow the use of our method in clinics, we have made all of the source codes for the analytics available via the internet, including the network-based analytics, E-GONet model, and data preprocessing methods [21].

Discussion

Using network analysis, we elucidated the relationships among antecedent behaviors prior to suicide attempts, where we identified the unique patterns associated with both lethal and nonlethal medical outcomes of suicide attempts. These findings allowed us to interpret the behaviors before lethal suicide attempts, thereby helping us to systematically investigate the interactions and connections among the behaviors that result in lethal suicide attempts. In particular, suicide attempts with lethal medical outcomes were associated with clear concepts regarding the likely fatality of the methods applied. In addition, behaviors, such as isolation at the time of suicide attempts, were strongly associated with the expected intervention time and the possibility of being discovered by other people. Among nonlethal suicide attempts, the suggested aim of suicide was a central factor among suicidal intent elements. Thus, lethal suicide attempts involved clear notions regarding the success of suicide, whereas nonlethal attempts were focused on the achievement of suicide attempts per se. In addition, prediction based on deep learning performed better after introducing the relationship signatures among the suicidal intent elements (60% increase in precision, 6% increase in the F1 score, and 8% increase in the AUPRG). Based on the analysis of feature contributions, we conclude that training of the relationships among SIS elements, especially isolation and the conception of a method's lethality, strongly ameliorated the deep learning performance. To the best of our knowledge, this is the first study to successfully discern the differences in mutual interactions among antecedent communicative behaviors prior to suicide attempts by those who had lethal and those who had nonlethal attempts, in which our novel method employed relational signatures to facilitate deep learning–based predictions.

In this study, we found that suicide attempts in individuals who had information about suicide methods and who anticipated fatality before attempting suicide had more lethal consequences. Our previous study showed that suicide methods are highly associated with subsequent suicide-related death [29]. Based on these results, we can infer that possessing information about suicide methods and their severity will affect suicide attempts and the consequent lethal outcome. Information about suicide methods can be found easily via the internet, and previous studies have shown that online searches for suicide-related terms are positively associated with intentional self-injury and death due to suicide [30,31]. In addition, we need to consider that suicide methods are subject to cultural differences. For example, suicide methods employed in the United States are predominantly related to firearms and the suicide rate is related to the gun possession rate by state [32]. By contrast, gun usage is very rare in South Korea because of the legal regulations related to gun possession [33]. However, the use of pesticides is fairly prevalent in suicide attempts in Korea, especially in rural areas and among elderly individuals who attempt suicide [17]. According to our results, we believe that restricting the accessibility of information regarding suicide methods is essential for suicide prevention, and cultural differences should be considered.

In previous studies, a machine learning algorithm trained with the longitudinal electronic health records of patients reliably predicted suicidal behavior [34] and actual suicide among US Army soldiers [35]. Linguistic-driven models that use the text in clinical notes have also been explored, but they lack sufficient accuracy (approximately 65%) [36]. In the future, machine learning based on medical big data may become a ubiquitous

component of clinical research and practice, which is a prospect that some find uncomfortable [37]. This study was based on three components comprising psychiatric physicians, data scientists, and a sophisticated computational infrastructure (KAT GPU Cluster System, Intel Xeon Ivy Bridge, 2.50 GHz 10 Cores; NVIDIA Tesla V100). However, the contributions of relationship features to the precise fatality predictions demonstrate that insights from physicians, including our hypothesis (ie, interactions among SIS elements were useful), as well as communication with the algorithm developer, are essential for innovative digital health development and precision medicine.

We have developed new approaches to investigate the characteristics of suicide attempts; however, this study had several limitations. First, the study sample did not represent the whole population of individuals who attempt suicide, as the 17 medical centers were located in specific urban areas of South Korea. The sample only included individuals who attempted suicide and came to the emergency centers [17]. In addition, the characteristics of suicide attempts differ among cultures. However, despite the limitations of the sample, the 1359 individuals who had suicide attempts comprised a large number of those who were assessed by a clinician shortly after their suicide attempts. The 17 medical centers were selected based on their enrollment in the National Emergency Department Information System, which is a government-managed nationwide registration system [38]. Lastly, E-GONet may have additional costs for learning and operating the deep learning model, as a deep learning model would require a more specialized facility with systems like a GPU system. Thus, cost-effectiveness and streamlined operation are the next milestones for deep learning–based approaches. For example, the world's best artificial intelligence model AlphaGo has a cost of US $35 million. This is much higher than the cost of a single human Go player per game.

As is usually noted in the clinical field, our data were relatively imbalanced (lethal, n=190; nonlethal, n=922). In order to appropriately analyze the data, we tried to use data resampling approaches. We applied an over-sampling method, an under-sampling method, and the synthetic minority over-sampling technique (SMOTE), but these methods did not improve the results. Therefore, we did not apply resampling approaches to our analysis. In addition, since resampling methods could not improve the results, it is expected that applying cost-sensitive loss functions will also not change the results.

In this study, we performed binary classification based on a lethality threshold (lethal >3, nonlethal ≤3). If we perform more fine-grained classification (ie, predict the exact C-SSRS grade), we could obtain more information for tailored care in clinics. However, as depicted in Table 1, the outcome of suicide attempts (C-SSRS grade) can be classified into six levels. Among the six levels, levels 5 and 6 involved very limited numbers of individuals who attempted suicide. Thus, we can only build a regression model for minor physical damage (ie, C-SSRS grades 1, 2, and 3) and severe damage (C-SSRS grade 4). However, because the number of individuals in C-SSRS grade 4 is limited compared with minor damage cases, the regression model for severe damage may not be well developed. Therefore, more fine-grained classification is not appropriate in this study. However, binary classification can provide clinically meaningful information. Regardless of our study design, the fine-grained identification of suicide attempts can be a guideline for further studies.

Two conclusions can be drawn regarding the originality of this study. First, effective management strategies can be provided for the care of individuals who attempt suicide, as individuals with lethal outcomes had expectations regarding the fatality of their suicide attempts and they made great preparations before their attempts so that they would not be found. Second, the enhanced performance of deep learning prediction shows that preprocessing the relationships in patient data using nonlinear transformation (ie, the interaction terms between SIS elements) can help the machine learning process understand the information embedded in clinical practice and employ it to make effective inferences.

The findings of this study may help public health officials and clinicians to identify the profiles of individuals at high risk of recurrent suicide attempts and may facilitate the development of efficient and effective suicide prevention programs.

Acknowledgments

This study was supported by the National Institute of General Medical Sciences of the National Institutes of Health (award number R01GM079719) and by the Korea Institute of Science and Technology Information (KISTI) (K-20-L02-C10-S01, K-17-L03-C02-S01, and P-18-SI-CT01-S01). The computational analysis was supported by the National Supercomputing Center, including resources and technology (K-18-L12-C08-S01, KAT GPU cluster system). YMA was supported by the Korean Ministry of Health (Korea National Suicide Survey [KNSS]).

Authors' Contributions

YMA and HP are both corresponding authors. BK and YK contributed equally to this manuscript. HP designed and wrote the paper; HP prepared the major figure, wrote the entire manuscript, designed the main algorithm, and analyzed the data; BK and YMA contributed to writing the manuscript and provided all of the data employed; YK conducted the deep learning process (E-GONet), feature contribution analysis, and background analysis; HKP and SJR contributed useful comments to writing the manuscript; YSK and BLL supported the project; and YMA and HP organized the overall project. All authors discussed the results and implications and commented on the manuscript at all stages.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structure of E-GONet, which is a convolutional neural network model.
[PNG File , 202 KB - medinform_v8i7e14500_app1.png]

Multimedia Appendix 2
Structure of E-GONet (without relationship features).
[PNG File , 138 KB - medinform_v8i7e14500_app2.png]

Multimedia Appendix 3
Importance of the utilized features for the training of E-GONet.
[PNG File , 160 KB - medinform_v8i7e14500_app3.png]

Multimedia Appendix 4
Detailed contribution of the Beck Suicide Intent Scale relationship features.
[PNG File , 153 KB - medinform_v8i7e14500_app4.png]

Multimedia Appendix 5
Precision-recall gain curves for random forest, linear regression, and E-GONet (with relationship features).
[PNG File , 149 KB - medinform_v8i7e14500_app5.png]

Multimedia Appendix 6
Performance of the area under the precision-recall gain curve.
[PNG File , 124 KB - medinform_v8i7e14500_app6.png]

References

1. Caine ED. Suicide and Attempted Suicide in the United States During the 21st Century. JAMA Psychiatry 2017 Nov 01;74(11):1087-1088. [doi: 10.1001/jamapsychiatry.2017.2524] [Medline: 28903162]
2. WISQARS Leading Causes of Death Reports. Centers for Disease Control and Prevention. URL: https://webappa.cdc.gov/sasweb/ncipc/leadcause.html [accessed 2018-12-28]
3. Lim D, Ha M, Song I. Trends in the leading causes of death in Korea, 1983-2012. J Korean Med Sci 2014 Dec;29(12):1597-1603 [FREE Full text] [doi: 10.3346/jkms.2014.29.12.1597] [Medline: 25469057]
4. O'Connor RC, Nock MK. The psychology of suicidal behaviour. Lancet Psychiatry 2014 Jun;1(1):73-85. [doi: 10.1016/S2215-0366(14)70222-6] [Medline: 26360404]
5. James K, Stewart D, Bowers L. Self-harm and attempted suicide within inpatient psychiatric services: a review of the literature. Int J Ment Health Nurs 2012 Aug;21(4):301-309. [doi: 10.1111/j.1447-0349.2011.00794.x] [Medline: 22340085]
6. Cavanagh JT, Carson AJ, Sharpe M, Lawrie SM. Psychological autopsy studies of suicide: a systematic review. Psychol Med 2003 Apr;33(3):395-405. [doi: 10.1017/s0033291702006943] [Medline: 12701661]
7. Mann JJ, Arango VA, Avenevoli S, Brent DA, Champagne FA, Clayton P, et al. Candidate endophenotypes for genetic studies of suicidal behavior. Biol Psychiatry 2009 Apr 01;65(7):556-563 [FREE Full text] [doi: 10.1016/j.biopsych.2008.11.021] [Medline: 19201395]
8. Beck AT, Beck R, Kovacs M. Classification of suicidal behaviors: I. Quantifying intent and medical lethality. Am J Psychiatry 1975 Mar;132(3):285-287. [doi: 10.1176/ajp.132.3.285] [Medline: 1115273]
9. Kumar CT, Mohan R, Ranjith G, Chandrasekaran R. Characteristics of high intent suicide attempters admitted to a general hospital. J Affect Disord 2006 Mar;91(1):77-81. [doi: 10.1016/j.jad.2005.12.028] [Medline: 16443283]
10. Rosen DH. The serious suicide attempt: epidemiological and follow-up study of 886 patients. Am J Psychiatry 1970 Dec;127(6):764-770. [doi: 10.1176/ajp.127.6.764] [Medline: 5482871]
11. Cramer AO, Waldorp LJ, van der Maas HL, Borsboom D. Comorbidity: a network perspective. Behav Brain Sci 2010 Jun;33(2-3):137-50; discussion 150. [doi: 10.1017/S0140525X09991567] [Medline: 20584369]
12. Borsboom D, Cramer AO. Network analysis: an integrative approach to the structure of psychopathology. Annu Rev Clin Psychol 2013;9:91-121. [doi: 10.1146/annurev-clinpsy-050212-185608] [Medline: 23537483]
13. Boschloo L, van Borkulo CD, Rhemtulla M, Keyes KM, Borsboom D, Schoevers RA. The Network Structure of Symptoms of the Diagnostic and Statistical Manual of Mental Disorders. PLoS One 2015;10(9):e0137621 [FREE Full text] [doi: 10.1371/journal.pone.0137621] [Medline: 26368008]

14. Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, et al. A network approach to psychopathology: new insights into clinical longitudinal data. PLoS One 2013;8(4):e60188 [FREE Full text] [doi: 10.1371/journal.pone.0060188] [Medline: 23593171]
15. van Borkulo C, Boschloo L, Borsboom D, Penninx BW, Waldorp LJ, Schoevers RA. Association of Symptom Network Structure With the Course of [corrected] Depression. JAMA Psychiatry 2015 Dec;72(12):1219-1226. [doi: 10.1001/jamapsychiatry.2015.2079] [Medline: 26561400]
16. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
17. Kim B, Ahn J, Cha B, Chung Y, Ha TH, Hong Jeong S, et al. Characteristics of methods of suicide attempts in Korea: Korea National Suicide Survey (KNSS). J Affect Disord 2015 Dec 01;188:218-225. [doi: 10.1016/j.jad.2015.08.050] [Medline: 26368946]
18. Pai D, Woo J, Son MH, Lee C. The Reliability and Validity of the Korean Version of Columbia-Suicide Severity Rating Scale in Alcohol Dependent Patients. J Korean Neuropsychiatr Assoc 2015;54(2):222. [doi: 10.4306/jknpa.2015.54.2.222]
19. Serrani Azcurra D. Psychometric validation of the Columbia-Suicide Severity rating scale in Spanish-speaking adolescents. Colomb Med (Cali) 2017 Dec 30;48(4):174-182 [FREE Full text] [doi: 10.25100/cm.v43i4.2294] [Medline: 29662259]
20. Beck RW, Morris JB, Beck AT. Cross-validation of the Suicidal Intent Scale. Psychol Rep 1974 Apr;34(2):445-446. [doi: 10.2466/pr0.1974.34.2.445] [Medline: 4820501]
21. Kim Y, Paik H. The suicide network source code. GitHub database. URL: https://github.com/hypaik/SuicideNetwork [accessed 2019-04-26]
22. Sambo F, Franzin A. bnstruct: Bayesian Network Structure Learning from Data with Missing Values. CRAN. URL: https://cran.r-project.org/web/packages/bnstruct/index.html [accessed 2018-08-31]
23. Bryant RA, Creamer M, O'Donnell M, Forbes D, McFarlane AC, Silove D, et al. Acute and Chronic Posttraumatic Stress Symptoms in the Emergence of Posttraumatic Stress Disorder: A Network Analysis. JAMA Psychiatry 2017 Feb 01;74(2):135-142. [doi: 10.1001/jamapsychiatry.2016.3470] [Medline: 28002832]
24. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003 Nov;13(11):2498-2504 [FREE Full text] [doi: 10.1101/gr.1239303] [Medline: 14597658]
25. TensorFlow. URL: http://www.tensorflow.org [accessed 2020-06-16]
26. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. 2018 Presented at: 6th International Conference on Learning Representations ICLR; 2018; Vancouver. [doi: 10.1007/978-3-030-28954-6_9]
27. Pflueger MO, Franke I, Graf M, Hachtel H. Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. BMC Psychiatry 2015 Mar 29;15:62 [FREE Full text] [doi: 10.1186/s12888-015-0447-4] [Medline: 25885691]
28. Flach PA, Kull M. Precision-Recall-Gain curves: PR analysis done right. Adv Neural Inf Process Syst 2015 [FREE Full text]
29. Kim B, Lee J, Kim E, Kim SH, Ha K, Kim YS, et al. Sex difference in risk period for completed suicide following prior attempts: Korea National Suicide Survey (KNSS). J Affect Disord 2018 Feb;227:861-868. [doi: 10.1016/j.jad.2017.11.013] [Medline: 29310206]
30. McCarthy MJ. Internet monitoring of suicide risk in the population. J Affect Disord 2010 May;122(3):277-279 [FREE Full text] [doi: 10.1016/j.jad.2009.08.015] [Medline: 19748681]
31. Yang AC, Tsai SJ, Huang NE, Peng CK. Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004-2009. J Affect Disord 2011 Jul;132(1-2):179-184. [doi: 10.1016/j.jad.2011.01.019] [Medline: 21371755]
32. Karp A. Completing the Count: Civilian firearms - Annexe online. In: Small Arms Survey 2007: Guns and the City; Chapter 2. Cambridge: Cambridge University Press; 2007.
33. Newton GD, Franklin EZ. Firearm Licensing: Permissive v Restrictive. In: Firearms & Violence in American Life: A staff report submitted to the National Commission on the Causes and Prevention of Violence. Washington, DC: US Government Printing Office; 1969.
34. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. Am J Psychiatry 2017 Feb 01;174(2):154-162. [doi: 10.1176/appi.ajp.2016.16010077] [Medline: 27609239]
35. Kessler RC, Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, Army STARRS Collaborators. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study To Assess Risk and rEsilience in Servicemembers (Army STARRS). JAMA Psychiatry 2015 Jan;72(1):49-57 [FREE Full text] [doi: 10.1001/jamapsychiatry.2014.1754] [Medline: 25390793]
36. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, et al. Predicting the risk of suicide by analyzing the text of clinical notes. PLoS One 2014;9(1):e85733 [FREE Full text] [doi: 10.1371/journal.pone.0085733] [Medline: 24489669]

37. Adkins DE. Machine Learning and Electronic Health Records: A Paradigm Shift. Am J Psychiatry 2017 Feb 01;174(2):93-94 [FREE Full text] [doi: 10.1176/appi.ajp.2016.16101169] [Medline: 28142275]
38. Cha W, Ahn K, Shin SD, Park J, Cho J. Emergency Department Crowding Disparity: a Nationwide Cross-Sectional Study. J Korean Med Sci 2016 Aug;31(8):1331-1336 [FREE Full text] [doi: 10.3346/jkms.2016.31.8.1331] [Medline: 27478347]

Abbreviations

AUPRG: area under the precision-recall gain curve
CNN: convolutional neural network
C-SSRS: Columbia Suicide Severity Rating Scale
ER: emergency room
GLASSO: graphical lasso
PPV: positive-predictive value
SIS: Beck Suicide Intent Scale

Edited by G Eysenbach; submitted 28.04.19; peer-reviewed by G Lim, S Kim; comments to author 12.08.19; revised version received 08.01.20; accepted 23.03.20; published 09.07.20.

Please cite as:
Kim B, Kim Y, Park CHK, Rhee SJ, Kim YS, Leventhal BL, Ahn YM, Paik H
Identifying the Medical Lethality of Suicide Attempts Using Network Analysis and Deep Learning: Nationwide Study
JMIR Med Inform 2020;8(7):e14500
URL: http://medinform.jmir.org/2020/7/e14500/
doi:10.2196/14500
PMID:32673253

Original Paper

Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study

Mark P Sendak[1], MPP, MD; William Ratliff[1], MBA; Dina Sarro[2], NP; Elizabeth Alderton[2], BSN; Joseph Futoma[3,4], PhD; Michael Gao[1], BS; Marshall Nichols[1], MS; Mike Revoir[1], BS; Faraz Yashar[3], BS; Corinne Miller[2], BSN; Kelly Kester[2], MSN; Sahil Sandhu[5], BS; Kristin Corey[1,6], MD; Nathan Brajer[1,6], MBA, MD; Christelle Tan[1,6], MD; Anthony Lin[1,6], MD; Tres Brown[7], BS; Susan Engelbosch[7], MBA; Kevin Anstrom[8], PhD; Madeleine Clare Elish[9], PhD; Katherine Heller[3,10], PhD; Rebecca Donohoe[11], MD; Jason Theiling[11], MD; Eric Poon[7,12], MPH, MD; Suresh Balu[1,6], MBA, MS; Armando Bedoya[7,13], MD; Cara O'Brien[12], MD

[1]Duke Institute for Health Innovation, Durham, NC, United States
[2]Duke University Hospital, Durham, NC, United States
[3]Department of Statistics, Duke University, Durham, NC, United States
[4]John A Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, United States
[5]Duke University, Durham, NC, United States
[6]Duke University School of Medicine, Durham, NC, United States
[7]Duke Health Technology Solutions, Durham, NC, United States
[8]Duke Clinical Research Institute, Durham, NC, United States
[9]Data & Society, New York, NY, United States
[10]Google, Mountain View, CA, United States
[11]Division of Emergency Medicine, Duke University School of Medicine, Durham, NC, United States
[12]Department of Medicine, Duke University School of Medicine, Durham, NC, United States
[13]Division of Pulmonary, Allergy, and Critical Care Medicine, Duke University School of Medicine, Durham, NC, United States

Corresponding Author:
Mark P Sendak, MPP, MD
Duke Institute for Health Innovation
200 Morris Street
3rd Floor
Durham, NC, 27701
United States
Phone: 1 919 684 3234
Email: mark.sendak@duke.edu

Abstract

Background: Successful integrations of machine learning into routine clinical care are exceedingly rare, and barriers to its adoption are poorly characterized in the literature.

Objective: This study aims to report a quality improvement effort to integrate a deep learning sepsis detection and management platform, Sepsis Watch, into routine clinical care.

Methods: In 2016, a multidisciplinary team consisting of statisticians, data scientists, data engineers, and clinicians was assembled by the leadership of an academic health system to radically improve the detection and treatment of sepsis. This report of the quality improvement effort follows the learning health system framework to describe the problem assessment, design, development, implementation, and evaluation plan of Sepsis Watch.

Results: Sepsis Watch was successfully integrated into routine clinical care and reshaped how local machine learning projects are executed. Frontline clinical staff were highly engaged in the design and development of the workflow, machine learning model, and application. Novel machine learning methods were developed to detect sepsis early, and implementation of the model required robust infrastructure. Significant investment was required to align stakeholders, develop trusting relationships, define roles and responsibilities, and to train frontline staff, leading to the establishment of 3 partnerships with internal and external research groups to evaluate Sepsis Watch.

Conclusions: Machine learning models are commonly developed to enhance clinical decision making, but successful integrations of machine learning into routine clinical care are rare. Although there is no playbook for integrating deep learning into clinical care, learnings from the Sepsis Watch integration can inform efforts to develop machine learning technologies at other health care delivery systems.

(JMIR Med Inform 2020;8(7):e15182) doi:10.2196/15182

KEYWORDS

machine learning; translational medicine; sepsis; innovation, organizational; change; deep learning

Introduction

Background

Technologies that digitize and harness massive amounts of data paired with the alignment of research and clinical care are transforming health care. Machine learning, a set of statistical methods optimized for prediction on new observations, is central to this transformation [1]. Although the translational pathway for prognostic models is well characterized, few machine learning models are externally validated or evaluated in clinical practice [2]. Isolated efforts demonstrating the clinical impact of previously validated technologies show the potential of machine learning in health care [3,4]. However, significant challenges remain for machine learning technologies to become fully embedded within standard operations of health care delivery systems [5].

Machine learning has been rapidly adopted in the biomedical sciences to enhance predictive, prognostic, and diagnostic methods. However, numerous technical and clinical barriers to its adoption persist. First, electronic health records (EHRs) often do not have native functionality to integrate complex machine learning models. Significant investment in infrastructure is required [6-8]. Second, even after a model is initially implemented, machine learning models can incur substantial ongoing maintenance costs [9-11]. Third, although some health systems do build and integrate home-grown machine learning solutions [12,13], that effort is often outsourced to research teams or technology vendors [5]. This divide between operations and model implementation and maintenance presents additional challenges, as "engineering ownership of the input signal is separate from the engineering ownership of the model that consumes it [10]." Finally, many models are not effectively integrated into clinical workflows in a fashion that improves clinical care or outcomes [14].

Sepsis Watch

Here, we delve into the details of how a health system integrated the first full-scale deep learning technology into routine clinical care. An innovation group spent over two years with partners across the organization to launch a deep learning solution, Sepsis Watch, on November 5, 2018. Sepsis Watch is a sepsis detection and management platform used by clinicians to improve compliance with recommended treatment guidelines for sepsis and thereby improve patient outcomes. Although Sepsis Watch is an instance of machine learning clinical decision support (CDS), deep learning systems do pose implementation challenges beyond traditional CDS, as detailed elsewhere [14-16]. In particular, new mechanisms of trust and accountability must be developed to ensure that the systems are safe and reliable [17,18]. In Table 1, we present the 8 steps required to integrate Sepsis Watch into routine care delivery successfully. We draw upon lessons from the learning health system framework and previously described best practices for responsible machine learning in health care [19,20]. The aim of this manuscript is to describe each step in detail and highlight learnings that can inform related efforts at other organizations.

Table 1. Steps for integrating machine learning into clinical care. The table includes definitions for the various steps and example tasks and deliverables during the step.

Step in the process	Definition	Example tasks and milestones
• Problem assessment	• Understand the root cause of the problem, the magnitude of the problem, where the problem is felt most acutely, who is best positioned to address the problem, and what changes need to occur to empower someone to address the problem	• Data analysis to understand the magnitude, setting, and timing of the problem • Observe frontline staff in clinical settings where the problem occurs • Interview a broad group of stakeholders to understand complexities in addressing the problem
• Internal and external scans of solutions and workflows	• Perform due diligence on internal and external tools that attempt to address the problem	• Evaluate technologies and workflows available through current information technology supplier relationships • Evaluate technologies on the market sold by external vendors • Interview internal stakeholders who have previously attempted to solve the problem
• Clinical workflow design	• Design clinical workflow that integrates new technology to address the problem	• Gather requirements from frontline staff and leadership • Iterate on workflow designs with frontline staff • Identify constraints (eg, time and effort) to ensure that the end user is able to use the technology effectively
• Model and infrastructure design	• Design machine learning model and accompanying infrastructure to ensure that the technology can effectively be integrated into clinical workflows	• Identify a set of input features used by the model to address the problem, making sure to incorporate clinical domain expertise and prior literature • Design infrastructure to support clinical decisions in a timely, actionable manner • Identify performance metrics and goals that are most important and relevant to stakeholders and end-users
• Clinical workflow application development	• Develop the clinical workflow application and integrations with other technologies	• Develop user interface and user experience • Integrate with electronic health record to access the required data at the required latency • Prototype workflow application with end users
• Model and infrastructure development	• Develop the machine learning model and infrastructure required to implement model, including integrations with other technologies	• Develop and validate the machine learning model on retrospective data • Validate the machine learning model and infrastructure on prospective *silent period* launch
• Implementation, change management, and governance	• Implement the machine learning model with accompanying education, communication, and governance to ensure accountability and successful adoption	• Establish a governance committee with agreed-upon tasks mission • Develop training material to ensure end users effectively use the new technology • Communicate broadly about the technology implementation and roles and responsibilities
• Evaluation plan and partnerships	• Prespecify evaluation plan, target goals, and safety and efficacy monitoring	• Develop internal and external partnerships to ensure rigorous evaluation • Register clinical trial

Methods

Problem Assessment

In October 2015, an interdisciplinary team of frontline clinicians at Duke Health proposed an innovation project to improve early detection of sepsis. With strong support from senior leadership, this project was launched in April 2016. The pilot project began at the academic flagship hospital, Duke University Hospital (DUH), and if the pilot yielded successful results, it would be expanded to 2 Duke Health community hospitals. Earlier attempts to implement CDS to improve timely detection and response to inpatient deterioration, including sepsis, caused significant alarm fatigue and did not improve clinical outcomes [21]. The Centers for Medicare and Medicaid Services (CMS) SEP-1 measure calculates compliance with 3-hour and 6-hour treatment bundles, and at the time the project began, in 2016, SEP-1 was progressing toward public reporting [22]. SEP-1 performance at DUH was poor, and clinical leaders found that although patients often received individual items of the sepsis bundle, follow-up items at 3 and 6 hours were often not completed.

Health system leaders wanted to reimagine how data and technology could be effectively utilized to both detect sepsis and coordinate care to ensure the completion of recommended

bundles. The team adopted computable sepsis criteria aligned with the CMS SEP-1 measure and quality improvement efforts at peer institutions, specified in Multimedia Appendix 1 [23,24]. Specific time windows to consider for each data element along with thresholds were decided upon by an interdisciplinary team of clinicians. Using these criteria, an analysis of DUH admissions data revealed that 55% and 68% of sepsis occurred within 12 hours and 24 hours, respectively, after presentation to the DUH emergency department (ED). Similarly, chart reviews of terminal hospitalizations involving sepsis found that over 70% of sepsis presented in the ED [25]. Hence, the initial clinical integration focused on improving sepsis detection and management among adults in the DUH ED. Table 2 displays characteristics of adult patients presenting between March and August 2018 to the site of the first integration, the DUH ED.

Internal and External Scans of Solutions and Workflows

In 2016, efforts to predict sepsis largely used the Medical Information Mart for Intensive Care [26] and restricted focus to just intensive care units (ICUs) [23,27,28]. At that time, there were successful reports of sepsis CDS algorithms [29], but there was no validated machine learning method to predict sepsis among adult patients presenting to the ED. A model specific to the ED predicted inpatient mortality among patients already meeting sepsis criteria [30]. The newly published quick Sequential Organ Failure Assessment (qSOFA) was recommended to identify patients at risk of poor outcomes because of sepsis [31]. However, qSOFA was not adopted by CMS for the SEP-1 core measure and does not accurately identify patients at risk of developing sepsis [24,32]. Considering the lack of an available, validated model to predict sepsis in the ED at that time, an interdisciplinary team of clinicians proposed to develop a novel machine learning model using local data.

The success of the innovation pilot depended on the rapid translation of model output to clinical action. The prior CDS implementation indicated that alert-fatigued frontline staff might not be the right personnel to receive alerts. The clinical team agreed that the rapid response team (RRT) nurses within DUH were best suited to triage patient alerts to manage sepsis proactively. RRTs significantly reduce time to medical resuscitation and escalation of care and, for sepsis specifically, improve the delivery of care bundles and outcomes [33-36]. The Sepsis Watch workflow needed to account for RRT nurses being mobile, caring for patients throughout the hospital, and needing to rapidly switch tasks to attend to urgent clinical duties.

Clinical Workflow Design

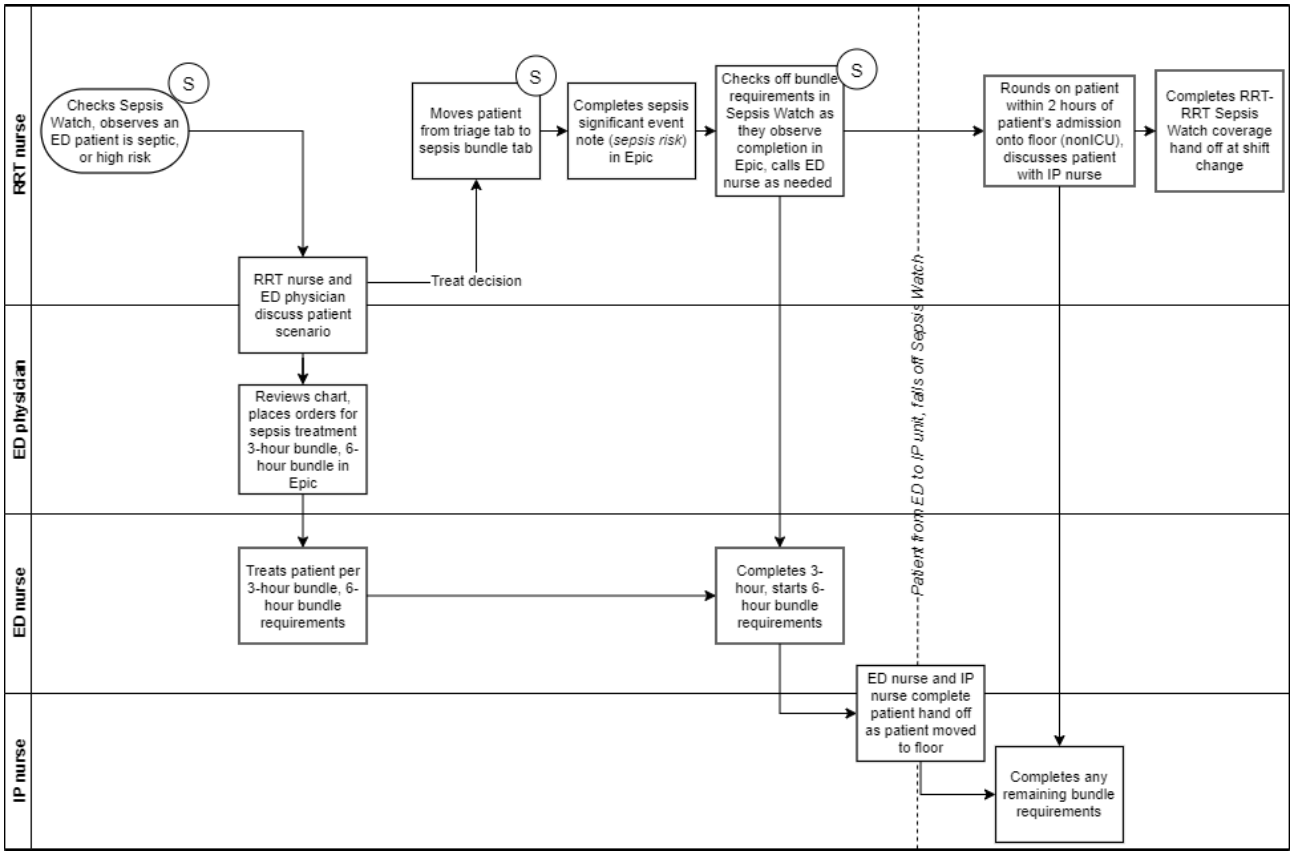
During the design phase, a transdisciplinary team of data scientists, statisticians, hospitalists, intensivists, ED clinicians, RRT nurses, and information technology leaders was assembled. The team designed the model and workflow concurrently and invested significant effort into gathering requirements from various stakeholders before writing code. Sepsis Watch was designed to be an overlay on top of existing clinical care within the ED, in contrast to the management of stroke and ST-elevation myocardial infarction, which require specialized teams to comanage patients alongside ED staff. For Sepsis Watch, all individual diagnostic and treatment actions are executed by the ED staff. Because of this distinction, the term *code sepsis* was avoided when describing Sepsis Watch. This workflow required a clear definition of roles and responsibilities across the RRT and ED clinical teams. Figure 1 presents the Sepsis Watch workflow. The RRT nurse triages patients at risk of sepsis using Sepsis Watch and communicates with ED clinicians about recommended treatment bundles. For patients who are confirmed to need treatment for sepsis, the RRT nurse enters a templated *significant event note* into the EHR. This note is meant to be a record of the RRT nurse evaluation and documentation for the admitting attending to carry out any remaining bundle items. The RRT nurse combined the use of Sepsis Watch with other clinical responsibilities across the hospital and communicated with ED clinicians telephonically. The Sepsis Watch user interface was carefully designed to accommodate tablet and mobile phone use.

Table 2. Cohort demographics for adults presenting to the Duke University Hospital emergency department between March 1, 2018, and August 31, 2018 (N=39,918).

Baseline characteristics of cohort	Values
Age (years), mean (SD)	50.41 (19.58)
Sex (male), n (%)	18,324 (45.90)
Admission source, n (%)	
Home or non–health care facility	34,892 (87.41)
Transfer from hospital	2887 (7.23)
Missing or other	2139 (5.36)
Admission type, n (%)	
Elective	5617 (14.07)
Emergency	30,099 (75.40)
Urgent	4160 (10.42)
Race, n (%)	
Black or African American	15,858 (39.73)
Caucasian or white	19,737 (49.44)
Missing or other	4323 (10.83)
Ethnicity, n (%)	
Not Hispanic/Latino	36,505 (91.45)
Hispanic/Latino	2352 (5.89)
Missing/other	1061 (2.66)
Comorbidities, n (%)	
Congestive heart failure	3349 (8.39)
Peripheral vascular disease	1685 (4.22)
Hypertension	11,934 (29.90)
Pulmonary circulation disorders	4063 (10.18)
Diabetes mellitus without chronic complications	2918 (7.31)
Solid tumor without metastasis	3949 (9.89)
Obesity	3225 (8.08)
Fluid and electrolyte disorders	5890 (14.76)
Anemia	3340 (8.37)
Depression	2733 (6.85)
Prior sepsis encounters in the past year, n (%)	
0	39,002 (97.71)
1	682 (1.71)
2 to 5	234 (0.59)
Septic, n (%)	2593 (6.50)
Emergency department	1377 53.10)
ICU[a]	468 (18.05)
General floor	602 (23.22)
Surgery	226 (8.72)
Overall rate of encounters that resulted in an admission, n (%)	18,620 (46.65)
Overall rate of ICU admission, n (%)	4668 (11.69)
Overall length of stay (hours), median (25% percentile, 75% percentile)	13.72 (5.11, 90.46)

[a]ICU: intensive care unit.

Figure 1. Sepsis Watch swimlane diagram.



Model and Infrastructure Design

The machine learning model was designed to detect sepsis early enough to provide clinicians time to confirm the diagnosis and complete CMS SEP-1 bundles. Model input features were both static (eg, prehospital patient comorbidities, patient demographics, and encounter details), and dynamic (eg, medication administrations, laboratory results, and vital measurements). Multimedia Appendix 1 lists all model features. The model was designed to perform well on adult patients who present to the ED from the time of ED triage through admission until time of death, discharge from the hospital, or admission to an ICU. The model was not trained to detect sepsis in the ICU setting. Patients admitted directly to surgery were excluded from model development. Similar to other sepsis prediction models [12], model inputs included both patient physiology (eg, vital sign measurements) and clinical interventions potentially prompted by suspicion of sepsis (eg, administration of antibiotics, measurement of serum lactate). In prior work, we demonstrated that inclusion of clinical interventions, such as indicators for whether or not a measurement or medication administration occurred in any given hour, improved model performance [37,38]. The practice of including clinical interventions as model inputs has been recommended for models built to be integrated into clinical practice [11].

Clinical leaders prioritized positive predictive value as a performance measure and were willing to trade-off model interpretability for performance gains. Model interpretability was low priority because of the many causes of sepsis, and treatment protocols are largely agnostic to cause. Sepsis Watch was designed to use the machine learning model to alert clinicians to evaluate patients further. Clinicians were instructed to put the model output into context with other relevant information to confirm or dismiss a sepsis diagnosis. The machine learning model did not drive clinical care in a standalone manner. The team worked closely with regulatory officials to ensure that Sepsis Watch qualified as CDS and was not a diagnostic medical device.

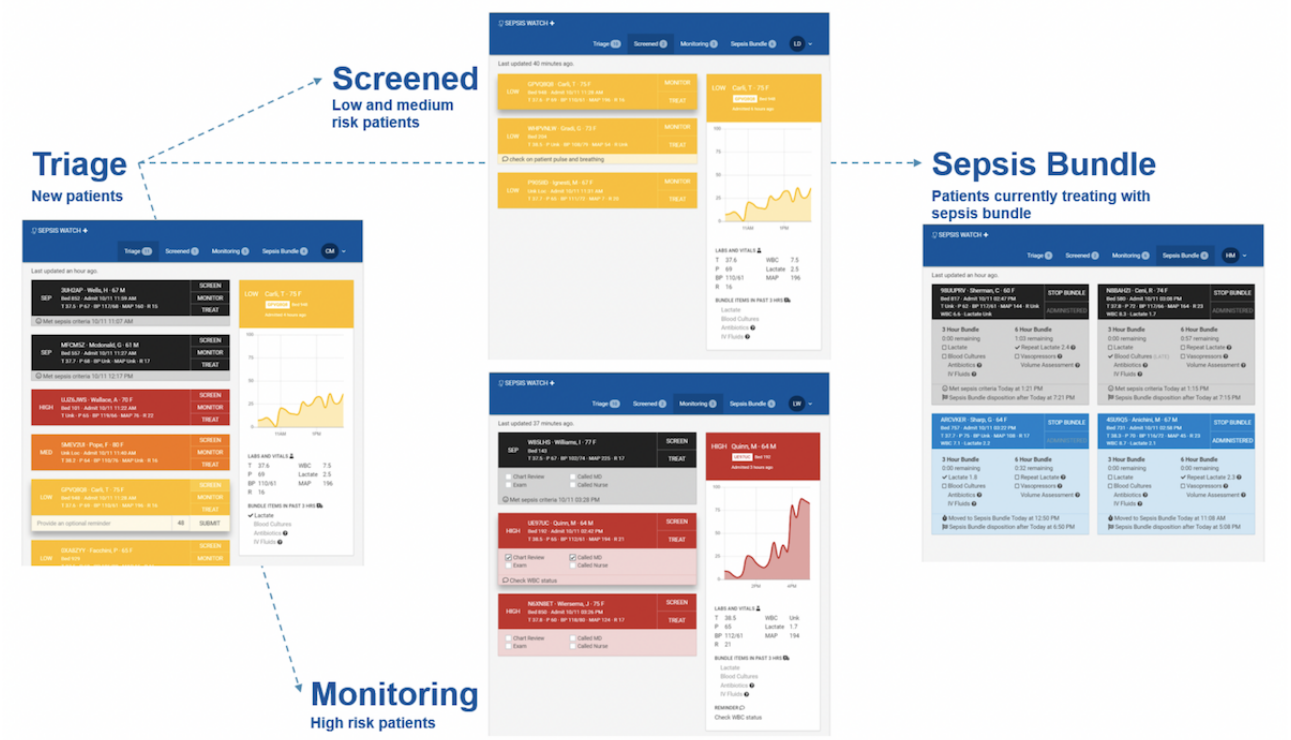
The team collaborated with technical and clinical stakeholders to define system requirements. The machine learning model needed to update the risk of sepsis for all patients every hour, whereas patients who met sepsis criteria needed to be identified every 5 minutes. Epic Web services needed to be built to allow Sepsis Watch to extract data from Epic every 5 minutes. The data are nurse-verified, and there are currently no interfaces to other monitors or data streams. The innovation pilot initially focused on an ED with approximately 200 visits per day and needed to be scalable to 1500 inpatient beds across the health system. The infrastructure was fully automated, fault-tolerant, parallelized, and run on on-premise computing infrastructure. During the 6-month pilot, the system uptime was 99.34%, with 1 instance of planned patching, 2 instances of the Web application being temporarily unavailable, and 1 instance of data not being updated. Sepsis Watch application code was to enhance portability and to scale across on-premise or cloud virtual environments while also improving reproducibility,

security, and ease of management. A database storing risk scores, time of sepsis, and information relevant to sepsis supported the Sepsis Watch Web application.

Clinical Workflow Application Development

After gathering requirements and iterating on designs of the workflow and model, development began in parallel on the Sepsis Watch Web application and a custom deep learning model. The Sepsis Watch Web application was developed in close collaboration with frontline staff. User interface designers repeatedly met with RRT nurses to iterate on functions, information, control, and visual components of the design. The first version contained 3 lists of patients: *Screening*, *Watchlist*, and *Treatment*. There was a 12-hour *snooze* state that prevented patients from being presented on the application. A second version removed the *snooze* state and included 4 lists of patients: *Triage*, *Screened*, *Monitoring*, and *Treatment*. This version ensured that all patients were visible at all times. The *Triage* page was the first point of entry for all patients presenting to the ED. Patients not requiring further evaluation were placed on the *Screened* page, whereas patients requiring further evaluation were placed on the *Monitoring* page. The *Treatment* page tracked completion of 3- and 6-hour sepsis bundle items for patients receiving treatment. Figure 2 displays screenshots of the application pages. Sepsis Watch was originally conceptualized as a dashboard to display model output; however, feedback and iterations led to the development of a highly interactive workflow management solution. Patients who met sepsis criteria were displayed in black colored cards, whereas patients at high risk of sepsis were displayed in red-colored cards. RRT nurses called ED physicians to discuss every patient with sepsis or at high risk of sepsis. No patient was placed on the *Treatment* page without independent review and confirmation by the attending physician.

Figure 2. Sepsis Watch user interface.



Model and Infrastructure Development

A novel machine learning method that combined multitask Gaussian processes (MGPs) and recurrent neural networks (RNNs) was developed to detect sepsis early [37,38]. RNNs are a type of deep learning configured to ingest time series data and are ideally suited to combine static and dynamic features of hospital encounters that vary in length [39,40]. The MGP learned distributions of continuous functions for each dynamic variable. Every hour, dynamic features sampled from the MGP were combined with static features and fed into the RNN to generate a risk of sepsis between 0 and 1. A separate set of scripts was optimized to run every 5 min to identify patients who meet sepsis criteria [41].

Our transdisciplinary team collaborated with Epic Systems to identify an optimal path for accessing clinical data in real time and integrating Sepsis Watch into the production system. A combination of off-the-shelf and custom-built Web services was utilized so that new patient information could be pulled every 5 min and stored in an external database. Given that the model ingests data from a variety of domains, 2 tools were developed to monitor Sepsis Watch. First, a Web service monitoring system and Web page was built to ensure that real-time integrations with Epic functioned appropriately. Second, a model monitoring system and Web page was built to display daily and weekly counts and mean values of model inputs and outputs.

Results

Implementation, Change Management, and Governance

A 3-month *silent period* was implemented before launch, during which Sepsis Watch first interacted with real-time data. A final round of data mapping was performed with clinical validation to reconcile changes in data formatting, followed by end-to-end testing of the model, data pipeline, user interface, and workflow. The first version of the model implemented the RNN without the MGP with minimal reduction in performance, reflecting practical trade-offs often made between model performance and engineering effort [42]. Thresholds were set to optimize positive predictive value and the number of alerts. Up to 4 high-risk alerts per hour were agreed upon as ideal volume for a single RRT nurse user. Clinical leaders reviewed 50 high-risk cases with a 72-hour delay to validate the threshold. Sepsis Watch accounts were created for clinical leaders to validate model output and test the workflow. Clinical leaders were instructed to contact inpatient teams if an observed patient needed immediate action. On an average day, about 14 patients met sepsis criteria, and about 7 patients were at high risk of sepsis.

Go-Live preparations focused on ensuring effective adoption and integration of Sepsis Watch into clinical care. Before Sepsis Watch, RRT nurses had minimal interaction with ED physicians. For Sepsis Watch to have the desired impact, these 2 roles needed to work closely together. With senior leadership support, regular touchpoints were prioritized to align partners around a unified vision of the workflow and potential impact. In the 4 weeks leading up to Go-Live, nearly a dozen hours per week were spent cultivating relationships and communication channels between roles. Weekly meetings brought together 1 to 2 leaders, each from the RRT nurse, ED nurse, ED physician, and inpatient hospitalist stakeholder groups. End-of-week updates were sent out every Friday, covering progress during the prior week and goals for the upcoming week to keep the team aligned. It was also during this time that the physicians and RRT nurses involved throughout the 2-year design and development process of Sepsis Watch served as crucial clinical champions promoting trust in the technology. In fact, the lead statisticians and developers had minimal interaction with frontline staff during the 4 weeks leading up to Go-Live. Clinicians with no formal information technology role within the health system promoted Sepsis Watch among their peers as a home-grown solution to an important problem within the hospital.

The next goal to drive adoption was to broadly communicate the change vision and empower action [43,44]. The team began in-person training for RRT nurses, emphasizing the urgent need to improve sepsis care in the ED and the opportunity to improve outcomes with Sepsis Watch. Although all RRT nurses worked in the critical care setting and were familiar with sepsis, training on the diagnostic criteria and treatment for sepsis was included to enhance awareness and understanding. The implementation team walked through the Sepsis Watch workflow with the RRT nurses in detail using a test version of the application populated with synthesized data. RRT nurses then interacted with the test version of the application themselves, iterating through the workflow steps, and asking questions to the implementation team. The in-person training ended with discussions of roles and responsibilities and the identification of various resources available to support frontline staff. Clinical nurse educators helped develop and distribute training content on an intranet webpage. Figure 3 shows a 1-page handout describing Sepsis Watch. This material was also communicated across clinical units through standing meetings and email listservs.

Sepsis Watch launched at 12 PM Eastern Daylight Time on November 5, 2018. The inaugural user was an RRT nurse who helped design the system. The ED medical director briefed the ED physicians to expect phone calls starting at noon. The RRT nurse was equipped with a tablet loaded with a link to the Sepsis Watch application, Epic's Canto app, the Sepsis Watch training homepage, contact information for all ED clinicians, a map of the ED, and a 2-min survey for submitting application and workflow feedback. The tablet and Sepsis Watch coverage were handed off at the end of each 12-hour shift. The Sepsis Watch Go-Live proceeded smoothly, and the application remained in continuous use by RRT nurses throughout the pilot.

The Sepsis Watch governance committee was created to monitor effectiveness and promote broad-based action. The committee included nursing, physician, and administrative leadership across the ED and inpatient wards. The committee's 4 primary goals were to (1) promote usage of the Sepsis Watch app, (2) provide comprehensive training and communication on the application and workflow, (3) develop a reporting method to track patient volume and bundle compliance, and (4) plan for postpilot sustainability. Table 3 lists the volume and bundle compliance metrics prioritized by the committee to include in weekly reports. These metrics provided clarity on compliance with specific bundle items, ensured that the volume of alerts was reasonable for a single RRT nurse user, and identified short-term wins to boost momentum. The implementation team sent weekly reports consisting of these metrics to frontline staff, including RRT nurse team members, and the Sepsis Watch governance committee.

Figure 3. Sepsis Watch training one-page overview.

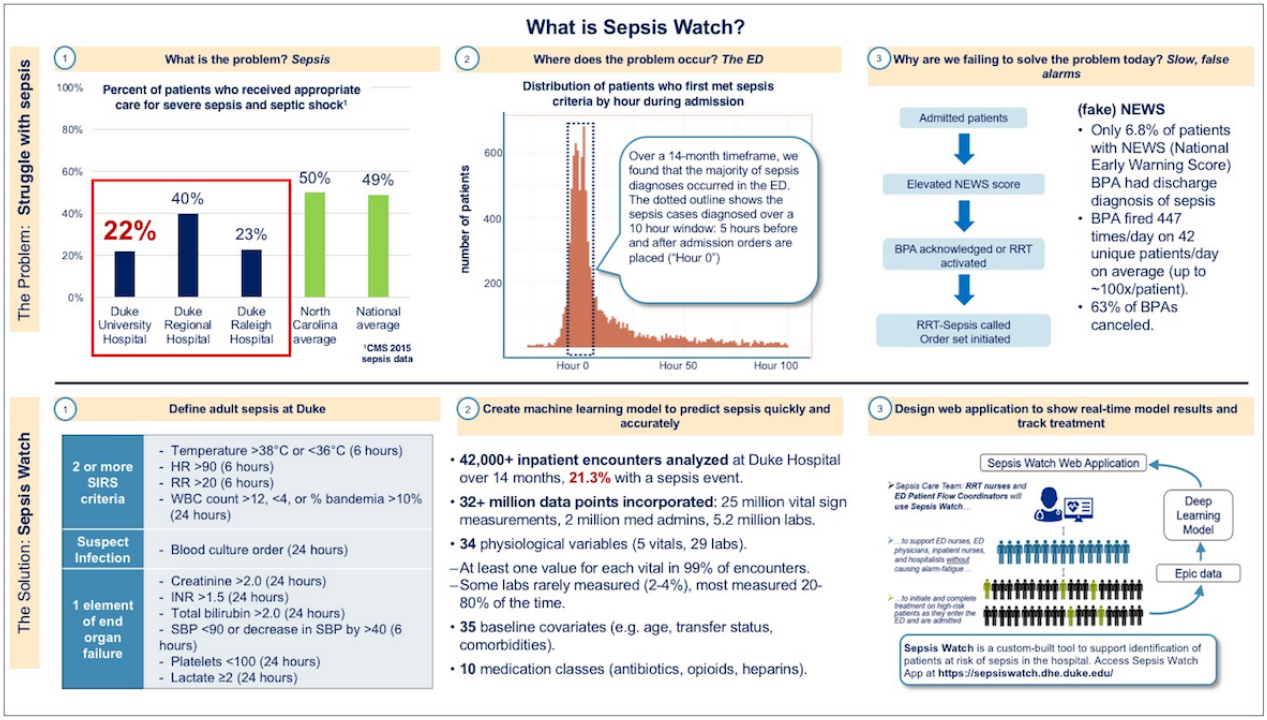


Table 3. Sepsis Watch governance weekly report metrics.

Metric types[a]	Metrics
Volume	Average number of new patients appearing on the Sepsis Watch Triage tab per day
Volume	Distribution of new patients appearing on the Sepsis Watch Triage tab, by hour of the day
Volume	Median length of time patient remained on the Sepsis Watch Triage tab before being moved to another tab
Volume	Average number of patients moved to the Sepsis Bundle Treatment tab per day
Bundle compliance	3-hour bundle compliance for patients moved to the Sepsis Watch Treatment tab (comprised of antibiotics, lactate, and blood culture 3-hour bundle components). Includes week-by-week performance
Bundle compliance	Antibiotics administration 3-hour compliance for patients moved to the Sepsis Watch Treatment tab. Includes week-by-week performance
Bundle compliance	Serum lactate collected 3-hour compliance for patients moved to the Sepsis Watch Treatment tab. Includes week-by-week performance
Bundle compliance	Blood culture collected 3-hour compliance for patients moved to the Sepsis Watch Treatment tab. Includes week-by-week performance

[a]Metrics were chosen by the Sepsis Watch governance committee to present data for 2 distinct patient cohorts: (1) patients who met Sepsis Watch sepsis criteria and (2) patients who were at high risk for meeting Sepsis Watch sepsis criteria as identified by the model.

Evaluation Plan and Partnerships

The Sepsis Watch evaluation consisted of continuous improvements to the workflow and user interface based on user feedback, 2 qualitative evaluations of how the technology impacted frontline clinicians, and a clinical and operational impact evaluation to demonstrate safety and efficacy. To enable continuous improvements, frontline staff regularly communicated feedback both indirectly, through a web-based survey that was bookmarked on the tablet, and directly to the team during regularly scheduled meetings or via email. These feedback loops were crucial to improving Sepsis Watch. All proposed changes and adaptations were prioritized and approved by the governance committee. For example, an early workflow change was to have the RRT nurse call the primary ED bedside nurse directly to ensure completion of sepsis treatments (eg, administration of antibiotics that are already ordered). Before this change, the RRT nurse was calling a charge nurse that managed ED intake triage, and the charge nurse was then expected to communicate with ED bedside nurses. We found that rather than centralizing information flow, the information needed to be communicated with the clinicians most directly involved in the care of patients who needed immediate action. Other changes related to improving communication channels included adopting *first call provider* functionality in the EHR to make explicit the covering physician for each patient that the RRT nurse needed to call and improving the layout of the phone number reference list for tablet use.

Similarly, a handful of changes were made to the user interface in the second version of the Sepsis Watch that was pushed out in January 2019. RRT nurses began using a standardized paper template to supplement Sepsis Watch on the tablet. The goal of the update was not to eliminate the need for a paper workflow supplement but to bring functionality that would further enhance efficiency into the application. For example, instead of comments being limited to patient transitions between lists, comments were enabled for all patients, and the character count was increased from 80 to 200. In addition, rather than only flagging sepsis bundle items that are complete (eg, blood culture collection and antibiotic administration), a flag was created for sepsis bundle items that are ordered (eg, blood culture ordered and antibiotic ordered). Before going live, the update was reviewed and approved by both frontline RRT nurses and the governance committee.

All new machine learning implementations have the potential for introducing inequality and bias that is not always clearly visible in numeric data [45-48]. Growing concern about such biases in health care highlighted the need for specialized evaluation of the Sepsis Watch [49-51]. Reflecting a commitment to rigorously study the outcomes and implications of integrating deep learning into clinical care, collaborations were established with 2 social science research institutes, Duke University's Social Science Research Institute (SSRI) and Data & Society Research Institute (D&S). Furthermore, 2 qualitative studies were designed to investigate the sociocultural dimensions of clinical integration. One evaluation, carried out with SSRI, focused on structured and semistructured interviews of ED physicians and RRT nurses, and the other evaluation, carried out with D&S, focused on observations of ED physicians and RRT nurses. Both studies analyzed Sepsis Watch in the context of organizational change management. These efforts aimed to identify adoption barriers and facilitators, and unintended social consequences and shifting clinical roles and responsibilities. Preliminary analysis of clinician's perceptions of evidence, trust, and authority in the early phase of development was completed, and several salient findings emerged [52]. First, building trust in the technology required much more than demonstrating model performance on a holdout and temporal validation set to clinicians. Stakeholders were looped in from the very beginning of the project, and it was important for the technology developers not to be or be seen as telling clinicians how to do their work. Second, the team identified the type and extent of evidence that was most salient to each stakeholder group. Although numbers and statistical trends were highlighted to hospital leaders, administrators, and managers, individual patient cases were important to frontline clinicians. This insight led to the development of patient-specific sepsis bundle reports that will go out to physicians and nurses involved in a patient case. Third, the team had to carefully navigate the lines of professional authority that physicians have toward the care of patients. Throughout the design, development, and implementation process, Sepsis Watch was described as a *tool* to support physicians and nurses in the ED, and the term *artificial intelligence* was not used in any communication or presentation.

The clinical impact will be evaluated (ClinicalTrials.gov ID: NCT03655626) after completion of the pilot. The primary outcome for this clinical trial is sepsis treatment bundle compliance. Secondary clinical outcomes include inpatient mortality, ICU requirement, and hospital and ED length of stay. Secondary process measures include time from ED presentation to meeting sepsis criteria and time from meeting sepsis criteria to completion of each bundle item. Table 4 presents baseline performance on a subset of clinical and process measures. The study includes balance measures to evaluate the overtreatment of patients at risk of sepsis. For example, the administration of antibiotics early in the clinical course of sepsis may not improve outcomes [53-55]. Several randomization schemes were considered to evaluate Sepsis Watch versus conventional treatment. However, expert clinicians can struggle when asked to complete a clinical task both with and without computer-aided support [56]. The intensive training of a small group of users also increased the risk of cross-group contamination. Ultimately, the single-site and operational nature of the innovation pilot made a prepost design most appropriate for this study. The data from this pilot can be used to recruit additional internal and external sites for a cluster-randomized trial to better characterize the causal relationship between Sepsis Watch and clinical outcomes [2].

Table 4. Baseline sepsis management performance at Duke University Hospital (n=1377).

Outcome measures	Baseline performance[a]
3-hour antibiotic compliance, n (%)	856 (62.16)
3-hour lactate compliance, n (%)	1064 (77.27)
3-hour blood culture compliance, n (%)	1237 (89.83)
3-hour antibiotic, lactate, and blood culture compliance	701 (50.91)
In-hospital mortality, n (%)	122 (8.86)
ICU[b] requirement, n (%)	491 (35.66)
Time from ED[c] arrival to meeting sepsis criteria (hours), median (25% percentile, 75% percentile)	1.93 (0.83, 5.08)
Length of stay in ED (hours), median (25% percentile, 75% percentile)	11.59 (9.87, 14.14)
Hospital length of stay overall (hours), median (25% percentile, 75% percentile)	125.42 (75.56, 215.18)

[a]Clinical and process measures for preimplementation cohort of adults who develop sepsis in the Duke University Hospital emergency department between March 1, 2018, and August 31, 2018.

[b]ICU: intensive care unit.

[c]ED: emergency department.

Discussion

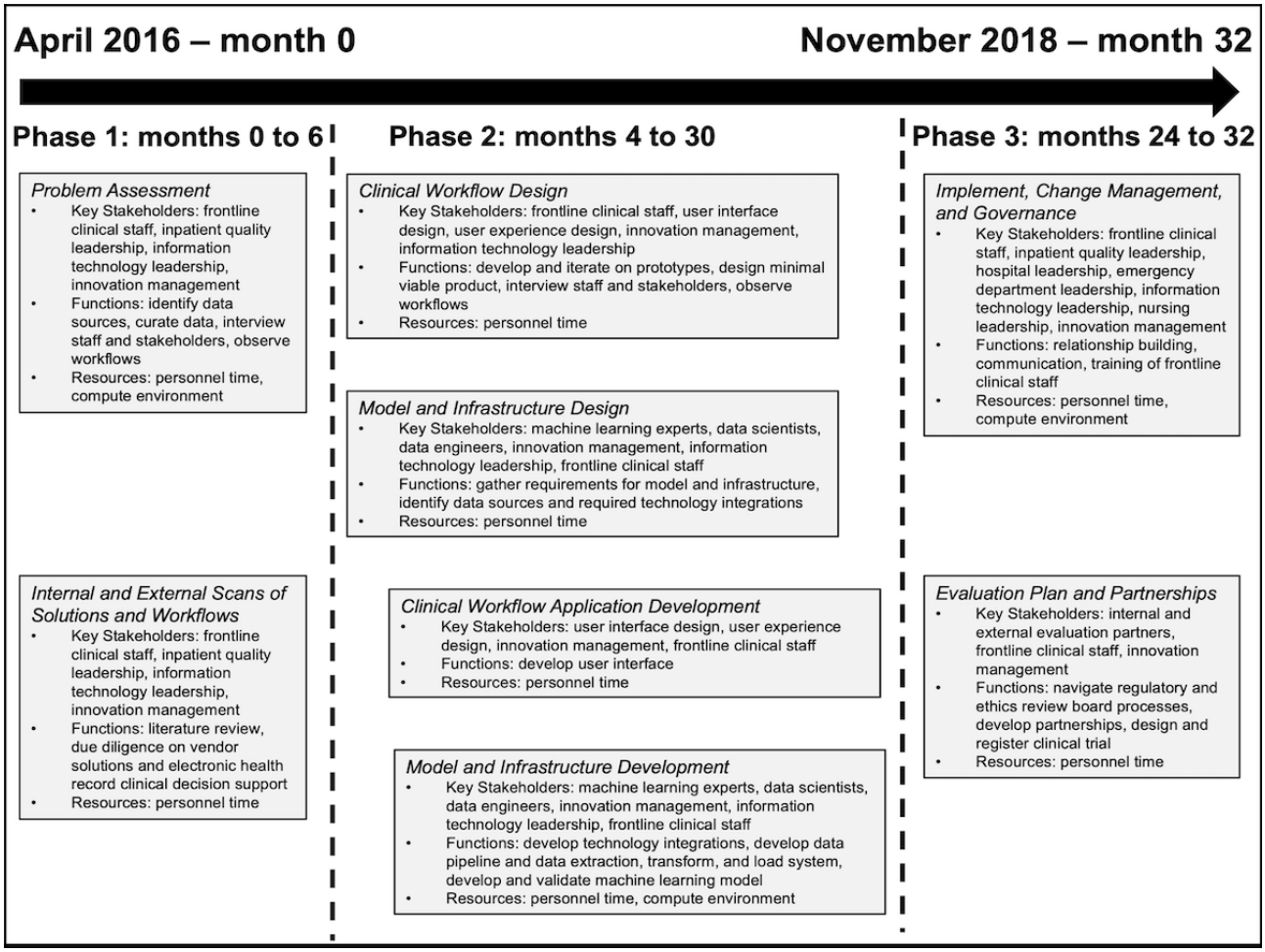
Principal Findings

What began as a 12-month innovation pilot to improve sepsis management became a multiyear groundbreaking effort that built capabilities, partnerships, and infrastructure with profound organizational impact. Figure 4 illustrates a timeline of the various steps described above, providing detail into the key stakeholders, functions, and resources involved in each step. Certain stakeholders, such as frontline clinical staff and innovation managers, were involved in every step of the process. Other stakeholders, such as hospital leaders and external research partners, were crucial to specific steps along the path. As the first integration of deep learning into routine care delivery, capabilities had to be developed across domains of expertise, requiring significant collaboration and cross-training. Physicians had to learn how to develop, use, and evaluate machine learning models. Information technology leaders had to learn how to integrate, support, and maintain machine learning models. Data scientists and machine learning experts had to learn about clinical data sources and sepsis. Although specialized skills were developed and will continue to be developed across the organization as additional projects are executed, we expect that close collaboration between clinicians, data scientists, data engineers, innovation managers, and information technology leaders will remain crucial.

The largest resource required for the successful translation of Sepsis Watch into routine clinical care was personnel time. Commodity compute infrastructure was used to support data analyses, model training, and model implementation. Personnel time was estimated for internal accounting purposes at about 8000 hours. Notably, many trainees were involved, including statistics graduate students, medical students, and clinical fellows, whose effort is not included in the estimate. This effort significantly exceeded the resources used for a prior effort to implement a linear regression model, estimated at US $220,000 [8]. However, the technology platform and capabilities built through the Sepsis Watch integration continues to create value for the organization. This infrastructure now supports additional applications of machine learning and accelerates research and quality improvement efforts across the organization. Other institutions that do not have in-house capabilities across technical and clinical domains may face additional costs and barriers. This reinforces previous findings that academic medical centers may be uniquely positioned to conduct translational machine learning research [5].

The steps presented in Figure 4 are not mutually exclusive and do not proceed in a neat, sequential manner. There is an overlap between the different phases and activities from an individual step may recur at a later phase of the project. For example, although the pilot began in the DUH ED, additional iterations of problem assessment were completed to better understand opportunities to improve sepsis care in other inpatient settings. Similarly, additional iterations and improvements to the user interface and workflow were made during the first few months of the pilot. Teams building and integrating machine learning technologies into routine clinical care should be prepared to iterate, maintain, and improve products throughout the product lifecycle.

Figure 4. Timeline of steps involved in translation of Sepsis Watch from problem identification to integration into routine clinical care.



Limitations

This study has several limitations. First, this is a single-center study, and the learnings and experience may not translate well to other settings. However, considering the lack of published evidence regarding the successful integration of machine learning into clinical care, initial case studies can be informative. Second, the integration of a deep learning model that changes clinical practice presents significant challenges with model updating. Feedback loops are created where Sepsis Watch may prompt clinical action for patients who do not ultimately develop sepsis [57]. Model retraining and updating will need to account for these feedback loops, which are an area of active research and will need to be explored in future work. Third, this study does not shed light on the predictors of sepsis and potential future directions of scientific inquiry. Fourth, this study does not present data on the clinical or economic impact of the integration of the Sepsis Watch. Analyses are underway and will be reported in future work. Fifth, this study does not demonstrate how well Sepsis Watch generalizes to care delivery settings beyond the ED. Future work will need to address generalizability both to external settings and other care units within the same hospital.

Conclusions

Despite the limitations, the successful integration of Sepsis Watch into routine clinical care signified a *crossing the chasm* journey for Duke Health [58]. Initially, a small number of visionary clinicians and administrators were eager to use emerging technology to address an important clinical problem. As the project progressed over two years, a broader group of stakeholders became aware of the potential impact of integrating machine learning into clinical care. A new request for applications was announced a month before the Sepsis Watch launch in November 2018, and the Duke Institute for Health Innovation received a record number of machine learning proposals, of which five machine learning proposals were ultimately selected by senior leadership and launched in April 2019 [59]. In June 2019, Sepsis Watch was disseminated to EDs at the two Duke Health community hospitals. Numerous challenges were encountered during the path to integration, but a focus on improving patient care moved Sepsis Watch from concept to design to production. There is no playbook for how to integrate machine learning into clinical care, and many more successful implementations are needed to develop best practices. Learnings from the Sepsis Watch integration have informed processes designed to improve the execution of machine learning projects within our health system. These learnings can provide direction to teams pursuing machine learning integrations into care elsewhere.

Acknowledgments

The authors would like to thank Drs Thomas Owens, William Fulkerson, Mary Klotman, Jeff Ferranti, Allan Kirk, Robert Califf, and Mary Ann Fuchs for operational and leadership support throughout the development of the Sepsis Watch. The authors would like to thank Drs Charles Gerardo and Allan Kirk for enthusiastically supporting the initial Sepsis Watch integration in the adult ED. Finally, the authors would like to thank Kevin Anstrom, Dan Mark, and Mary Ann Fuchs for advising on the Sepsis Watch clinical study. The development, validation, and integration of Sepsis Watch into clinical care was fully funded by the Duke Health, with significant contributions by the Duke Institute for Health Innovation and Duke Health Technology Solutions. Data & Society's qualitative evaluation was supported by funding from Luminate, the John D and Catherine T MacArthur Foundation, and the Ethics and Governance of Artificial Intelligence Fund. Funders were not involved in the preparation, review, or submission of the manuscript.

Conflicts of Interest

MS, WR, JF, MG, MN, MR, NB, AL, KH, AB, and CO are named inventors of the Sepsis Watch deep learning model, which was licensed from Duke University by Cohere Med, Inc. These authors do not hold any equity in Cohere Med, Inc. No other authors have relevant financial disclosures.

Multimedia Appendix 1
Supplement to the manuscript including two tables.
[DOCX File , 15 KB - medinform_v8i7e15182_app1.docx]

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, PROGRESS Group. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381 [FREE Full text] [doi: 10.1371/journal.pmed.1001381] [Medline: 23393430]
3. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res 2017;4(1):e000234 [FREE Full text] [doi: 10.1136/bmjresp-2017-000234] [Medline: 29435343]
4. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39 [FREE Full text] [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
5. Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine learning in health care: a critical appraisal of challenges and opportunities. EGEMS (Wash DC) 2019 Jan 24;7(1):1 [FREE Full text] [doi: 10.5334/egems.287] [Medline: 30705919]
6. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, et al. Health care and precision medicine research: analysis of a scalable data science platform. J Med Internet Res 2019 Apr 9;21(4):e13043 [FREE Full text] [doi: 10.2196/13043] [Medline: 30964441]
7. Corey K, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. PLoS Med 2018 Nov;15(11):e1002701 [FREE Full text] [doi: 10.1371/journal.pmed.1002701] [Medline: 30481172]
8. Sendak M, Balu S, Schulman K. Barriers to achieving economies of scale in analysis of EHR data. A cautionary tale. Appl Clin Inform 2017 Aug 9;8(3):826-831 [FREE Full text] [doi: 10.4338/ACI-2017-03-CR-0046] [Medline: 28837212]
9. Sculley D, Holt G, Golovin D. Hidden Technical Debt in Machine Learning Systems. NIPS Proceedings. 2014. URL: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf [accessed 2020-06-16]
10. Sculley D, Holt G, Golovin D. Machine Learning: The High Interest Credit Card of Technical Debt. Google Research. 2014. URL: https://research.google/pubs/pub43146/ [accessed 2020-06-16]
11. Lenert M, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. J Am Med Inform Assoc 2019 Dec 1;26(12):1645-1650. [doi: 10.1093/jamia/ocz145] [Medline: 31504588]
12. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr;46(4):547-553 [FREE Full text] [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
13. Giannini H, Ginestra J, Chivers C, Draugelis M, Hanish A, Schweickert WD, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Crit Care Med 2019 Nov;47(11):1485-1492. [doi: 10.1097/CCM.0000000000003891] [Medline: 31389839]

14. Emanuel E, Wachter RM. Artificial intelligence in health care: will the value match the hype? J Am Med Assoc 2019 Jun 18;321(23):2281-2282. [doi: 10.1001/jama.2019.4914] [Medline: 31107500]
15. Jamieson T, Goldfarb A. Clinical considerations when applying machine learning to decision-support tasks versus automation. BMJ Qual Saf 2019 Oct;28(10):778-781. [doi: 10.1136/bmjqs-2019-009514] [Medline: 31147420]
16. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res 2019 Jul 10;21(7):e13659 [FREE Full text] [doi: 10.2196/13659] [Medline: 31293245]
17. Saria S, Subbaswamy A. Tutorial: Safe and Reliable Machine Learning. In: ACM Conference on Fairness, Accountability, and Transparency. 2019 Jan Presented at: FAT'19; January 29-31, 2019; Atlanta, USA URL: https://arxiv.org/pdf/1904.07204.pdf
18. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. J Am Med Assoc 2019 Jul 15:- epub ahead of print. [doi: 10.1001/jama.2018.20563] [Medline: 31305873]
19. Greene SM, Reid RJ, Larson EB. Implementing the learning health system: from concept to action. Ann Intern Med 2012 Aug 7;157(3):207-210. [doi: 10.7326/0003-4819-157-3-201208070-00012] [Medline: 22868839]
20. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019 Sep;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
21. Bedoya AD, Clement ME, Phelan M, Steorts RC, O'Brien C, Goldstein BA. Minimal impact of implemented early warning score and best practice alert for patient deterioration. Crit Care Med 2019 Jan;47(1):49-55 [FREE Full text] [doi: 10.1097/CCM.0000000000003439] [Medline: 30247239]
22. Rhodes A, Evans L, Alhazzani W, Levy MM, Antonelli M, Ferrer R, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. Intensive Care Med 2017 Mar;43(3):304-377. [doi: 10.1007/s00134-017-4683-6] [Medline: 28101605]
23. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 5;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
24. Amland RC, Sutariya BB. Quick sequential [sepsis-related] organ failure assessment (qSOFA) and St John sepsis surveillance agent to detect patients at risk of sepsis: an observational cohort study. Am J Med Qual 2018;33(1):50-57 [FREE Full text] [doi: 10.1177/1062860617692034] [Medline: 28693336]
25. Rhee C, Jones TM, Hamad Y, Pande A, Varon J, O'Brien C, Centers for Disease Control and Prevention (CDC) Prevention Epicenters Program. Prevalence, underlying causes, and preventability of sepsis-associated mortality in us acute care hospitals. JAMA Netw Open 2019 Feb 1;2(2):e187571 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.7571] [Medline: 30768188]
26. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
27. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. Comput Biol Med 2016 Jul 1;74:69-73. [doi: 10.1016/j.compbiomed.2016.05.003] [Medline: 27208704]
28. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 2016 Sep 30;4(3):e28 [FREE Full text] [doi: 10.2196/medinform.5909] [Medline: 27694098]
29. Brandt B, Gartner AB, Moncure M, Cannon CM, Carlton E, Cleek C, et al. Identifying severe sepsis via electronic surveillance. Am J Med Qual 2015;30(6):559-565. [doi: 10.1177/1062860614541291] [Medline: 24970280]
30. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. Acad Emerg Med 2016 Mar;23(3):269-278 [FREE Full text] [doi: 10.1111/acem.12876] [Medline: 26679719]
31. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). J Am Med Assoc 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
32. Askim A, Moser F, Gustad LT, Stene H, Gundersen M, Åsvold BO, et al. Poor performance of quick-SOFA (qSOFA) score in predicting severe sepsis and mortality - a prospective study of patients admitted with infection to the emergency department. Scand J Trauma Resusc Emerg Med 2017 Jun 9;25(1):56 [FREE Full text] [doi: 10.1186/s13049-017-0399-4] [Medline: 28599661]
33. Ju T, Al-Mashat M, Rivas L, Sarani B. Sepsis rapid response teams. Crit Care Clin 2018 Apr;34(2):253-258. [doi: 10.1016/j.ccc.2017.12.004] [Medline: 29482904]
34. Sebat F, Musthafa AA, Johnson D, Kramer AA, Shoffner D, Eliason M, et al. Effect of a rapid response system for patients in shock on time to treatment and mortality during 5 years. Crit Care Med 2007 Nov;35(11):2568-2575. [doi: 10.1097/01.CCM.0000287593.54658.89] [Medline: 17901831]
35. Umscheid CA, Betesh J, van Zandbergen C, Hanish A, Tait G, Mikkelsen ME, et al. Development, implementation, and impact of an automated early warning and response system for sepsis. J Hosp Med 2015 Jan;10(1):26-31 [FREE Full text] [doi: 10.1002/jhm.2259] [Medline: 25263548]

36. Guirgis FW, Jones L, Esma R, Weiss A, McCurdy K, Ferreira J, et al. Managing sepsis: electronic recognition, rapid response teams, and standardized care save lives. J Crit Care 2017 Aug;40:296-302 [FREE Full text] [doi: 10.1016/j.jcrc.2017.04.005] [Medline: 28412015]
37. Futoma J, Hariharan S, Heller K. Learning to Detect Sepsis With a Multitask Gaussian Process RNN Classifier. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. 2017 Jun Presented at: ICM'17; August 6-11, 2017; Sydney, Australia URL: https://dl.acm.org/doi/10.5555/3305381.3305503 [doi: 10.5555/3305381.3305503]
38. Futoma J, Hariharan S, Sendak M, Brajer N, Clement M, Bedoya A, et al. An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. In: Proceedings of Machine Learning for Healthcare. 2017 Presented at: MLH'17; August 19-20, 2017; Los Angeles, CA, USA URL: https://arxiv.org/abs/1708.05894
39. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018 Dec;6(12):905-914. [doi: 10.1016/S2213-2600(18)30300-X] [Medline: 30274956]
40. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018 Apr 17;8(1):6085 [FREE Full text] [doi: 10.1038/s41598-018-24271-9] [Medline: 29666385]
41. Sendak M. constellation: Identify Event Sequences Using Time Series Joins. The Comprehensive R Archive Network. 2018. URL: https://cran.r-project.org/web/packages/constellation/index.html [accessed 2019-06-25]
42. Johnston C. Netflix Never Used Its $1 Million Algorithm Due To Engineering Costs. Wired. 2012. URL: https://www.wired.com/2012/04/netflix-prize-costs/ [accessed 2019-06-25]
43. Kotter J. Leading Change: Why Transformation Efforts Fail. Harvard Business Review - Ideas and Advice for Leaders. 1995. URL: https://hbr.org/1995/05/leading-change-why-transformation-efforts-fail-2 [accessed 2020-06-16]
44. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 7;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
45. Eubanks V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York, USA: St Martin's Press; 2018.
46. Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks. ProPublica. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [accessed 2019-06-25]
47. Bogen M, Rieke A. Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias. Upturn. 2018. URL: https://www.upturn.org/reports/2018/hiring-algorithms/ [accessed 2019-06-25]
48. Barocas S, Selbst AD. Big data's disparate impact. Cal L Rev 2016 Jun;104(3):671-732 [FREE Full text]
49. Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. JAMA Intern Med 2019 Mar 1;179(3):293-294. [doi: 10.1001/jamainternmed.2018.7117] [Medline: 30556825]
50. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 1;178(11):1544-1547 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
51. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. J Am Med Assoc 2017 Aug 8;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
52. Elish MC. The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care. In: Ethnographic Praxis in Industry Conference. 2018 Presented at: EPIC'18; October 9-12, 2018; Honolulu, HI. [doi: 10.1111/1559-8918.2018.01213]
53. Alam N, Oskam E, Stassen PM, Exter PV, van de Ven PM, Haak HR, PHANTASi Trial Investigatorsthe ORCA (Onderzoeks Consortium Acute Geneeskunde) Research Consortium the Netherlands. Prehospital antibiotics in the ambulance for sepsis: a multicentre, open label, randomised trial. Lancet Respir Med 2018 Jan;6(1):40-50. [doi: 10.1016/S2213-2600(17)30469-1] [Medline: 29196046]
54. Klompas M, Calandra T, Singer M. Antibiotics for sepsis-finding the equilibrium. J Am Med Assoc 2018 Oct 9;320(14):1433-1434. [doi: 10.1001/jama.2018.12179] [Medline: 30242350]
55. Mi MY, Klompas M, Evans L. Early administration of antibiotics for suspected sepsis. N Engl J Med 2019 Feb 7;380(6):593-596. [doi: 10.1056/NEJMclde1809210] [Medline: 30726686]
56. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL, Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015 Nov;175(11):1828-1837 [FREE Full text] [doi: 10.1001/jamainternmed.2015.5231] [Medline: 26414882]
57. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 18;169(12):866-872 [FREE Full text] [doi: 10.7326/M18-1990] [Medline: 30508424]
58. Moore G. Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers. New York, USA: Harper Business; 2006.
59. DIHI Announces 2019 RFA Innovation Awards. Duke OLV – The Office of Licensing & Ventures. 2019. URL: https://olv.duke.edu/news/dihi-announces-2019-rfa-innovation-awards/ [accessed 2019-06-25]

Abbreviations

CDS: clinical decision support
CMS: Centers for Medicare & Medicaid Services
D&S: Data & Society Research Institute
DUH: Duke University Hospital
ED: emergency department
EHR: electronic health record
ICU: intensive care unit
MGP: multitask Gaussian process
qSOFA: quick Sequential Organ Failure Assessment
RNN: recurrent neural network
RRT: rapid response team
SSRI: Social Sciences Research Institute

Edited by G Eysenbach; submitted 26.06.19; peer-reviewed by V Liu, M Banf, F Lanfranchi, S Nemati; comments to author 30.09.19; revised version received 23.11.19; accepted 31.12.19; published 15.07.20.

Please cite as:
Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, Nichols M, Revoir M, Yashar F, Miller C, Kester K, Sandhu S, Corey K, Brajer N, Tan C, Lin A, Brown T, Engelbosch S, Anstrom K, Elish MC, Heller K, Donohoe R, Theiling J, Poon E, Balu S, Bedoya A, O'Brien C
Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study
JMIR Med Inform 2020;8(7):e15182
URL: http://medinform.jmir.org/2020/7/e15182/
doi:10.2196/15182
PMID:32673244

Original Paper

Applicability of an Automated Model and Parameter Selection in the Prediction of Screening-Level PTSD in Danish Soldiers Following Deployment: Development Study of Transferable Predictive Models Using Automated Machine Learning

Karen-Inge Karstoft[1,2], PhD; Ioannis Tsamardinos[3,4], PhD; Kasper Eskelund[1,5], PhD; Søren Bo Andersen[1], PhD; Lars Ravnborg Nissen[1], MD

[1]Research and Knowledge Centre, The Danish Veterans Centre, Ringsted, Denmark
[2]Department of Psychology, University of Copenhagen, Copenhagen, Denmark
[3]Department of Computer Science, University of Crete, Heraklion, Crete, Greece
[4]Gnosis Data Analysis PC, Heraklion, Greece
[5]Department of Military Psychology, The Danish Veterans Centre, Copenhagen, Denmark

Corresponding Author:
Karen-Inge Karstoft, PhD
Research and Knowledge Centre
The Danish Veterans Centre
Garnisonen 1
Ringsted,
Denmark
Phone: 45 61671619
Email: kareningekarstoft@gmail.com

Abstract

Background: Posttraumatic stress disorder (PTSD) is a relatively common consequence of deployment to war zones. Early postdeployment screening with the aim of identifying those at risk for PTSD in the years following deployment will help deliver interventions to those in need but have so far proved unsuccessful.

Objective: This study aimed to test the applicability of automated model selection and the ability of automated machine learning prediction models to transfer across cohorts and predict screening-level PTSD 2.5 years and 6.5 years after deployment.

Methods: Automated machine learning was applied to data routinely collected 6-8 months after return from deployment from 3 different cohorts of Danish soldiers deployed to Afghanistan in 2009 (cohort 1, N=287 or N=261 depending on the timing of the outcome assessment), 2010 (cohort 2, N=352), and 2013 (cohort 3, N=232).

Results: Models transferred well between cohorts. For screening-level PTSD 2.5 and 6.5 years after deployment, random forest models provided the highest accuracy as measured by area under the receiver operating characteristic curve (AUC): 2.5 years, AUC=0.77, 95% CI 0.71-0.83; 6.5 years, AUC=0.78, 95% CI 0.73-0.83. Linear models performed equally well. Military rank, hyperarousal symptoms, and total level of PTSD symptoms were highly predictive.

Conclusions: Automated machine learning provided validated models that can be readily implemented in future deployment cohorts in the Danish Defense with the aim of targeting postdeployment support interventions to those at highest risk for developing PTSD, provided the cohorts are deployed on similar missions.

(JMIR Med Inform 2020;8(7):e17119) doi:10.2196/17119

KEYWORDS

decision support; machine learning; mental health; PTSD; military; screening

Introduction

Posttraumatic stress disorder (PTSD) is a relatively common problem following exposure to trauma [1]. Following deployment to war zones, PTSD or symptoms thereof are seen in a significant percentage of soldiers. A recent review found average PTSD rates of 12.9% (95% CI 11.3%-14.4%) for military personnel deployed to Iraq and 7.1% (95% CI 4.6%-9.6%) for personnel deployed to Afghanistan among military personnel from the United States, the United Kingdom, and Canada [2]. Among Danish soldiers, cohort studies have found that approximately 10% experience severe symptoms of PTSD 2.5 years after returning from deployment to Afghanistan [3]. With a total of 9949 Danish soldiers deployed to Afghanistan as of December 31, 2018 and 33,131 deployed to different combat zones including Afghanistan, Iraq, and the Balkans, this poses a significant public health problem.

Preventing large-scale bouts of PTSD and the derivative effects among previously deployed soldiers calls for reliable screening tools that can be applied early relative to deployment [4]. However, previous efforts at mental health screening among soldiers before deployment and shortly after returning have so far proved futile [5]. As such, research efforts to date have not enabled primary prevention, where highly vulnerable individuals with a high risk of developing PTSD following deployment are not deployed, or secondary prevention, where early treatment is offered to those in need, best preventing them from developing severe or chronic PTSD [6].

One reason for this lack of success might be the use of traditional statistics when investigating and integrating risk factors into predictive models of postdeployment PTSD [7]. Such models may not be able to account for the multidimensionality and nonlinearity of interacting risk factors for PTSD [8]; as such, they fail to provide an accurate prediction. In recent years, methods of machine learning (ML) have made their way into the literature on trauma reactions and in psychiatry more generally, with the primary aim of early prediction of psychological problems or psychiatric diagnoses such as PTSD [9,10]. ML is a broad term covering computational methods that work by learning from data with the aim of building models that are able to recognize patterns, distinguish between categories, or predict the level or degree of some trait or characteristic [11]. In the specific case of supervised ML, an algorithm is trained in relation to some outcome of interest, with the aim of categorizing individuals as belonging to one or another predefined category [12].

A recent review described 15 studies applying ML methods to predict PTSD or categorize individuals as PTSD cases or noncases [8] and found that, in general, ML prediction models applied to the domain of PTSD prediction reveal promising results. Most of the papers included in the review were cross-sectional (both predictor values and the outcome were measured at the same time point); hence, they were not predictive, but diagnostic, of PTSD status. In general, these cross-sectional studies classified PTSD with very high accuracy (area under the receiver operating characteristic [ROC] curve [AUC] ranging from 0.79 to 0.97). Another group of papers in the review predicted PTSD at a follow-up of <1 year; in general, the prediction accuracies of these studies are respectable. For example, Saxe et al [13] and Rosellini et al [14] both aimed to predict PTSD at 3 months after trauma or release from the hospital, respectively, and did so with high accuracy (AUC=0.79). Finally, three papers in the review predicted PTSD at a follow-up of >1 year, 2 with a follow-up at 15 months [15,16] and one with follow-up at 2.5 years [17]. All three studies achieved acceptable to high accuracy (AUCs ranging from 0.75 to 0.88). While these results are promising, it is clear that more ML prediction studies with longer follow-ups are needed to test the applicability of such methods in practice. Further, focus on future efforts within ML prediction of PTSD should be on how the trained and tested algorithms can be implemented in clinical contexts and for screening purposes.

In the military context, Kessler and colleagues [18] applied ML techniques to predict suicide after hospitalization with psychiatric diagnoses in service members, whereas Rosellini and colleagues [19] used ML to predict postdeployment psychiatric disorder symptoms and interpersonal violence during deployment based on predeployment characteristics. Both studies found that ML provides relatively accurate prediction of the targeted outcomes, with the best performing predictive models in the study by Rosellini et al [19] significantly outperforming logistic regression models. In an earlier study by our own group, we applied a specific ML algorithm, namely support vector machine (SVM), in a cohort study with Danish soldiers deployed to Afghanistan with the aim of predicting PTSD symptomatology 2.5 years after deployment based on predeployment and early postdeployment characteristics [17]. Briefly, we found that long-term posttraumatic stress could be predicted with good accuracy by predeployment indicators (AUC=0.84) and by predeployment indicators combined with indicators collected immediately postdeployment (AUC=0.88).

These studies applied ML to predict PTSD using a variety of ML methods, each requiring expertise and resources for the selection of appropriate algorithms and tuning of hyperparameters. Automated machine learning (AutoML) is a quickly rising subfield of ML promising to ease the application while ensuring correct and optimal utilization of ML methods [20]. Here, we use and test AutoML as implemented in the Just Add Data Bio (JADBio) tool [21]. In brief, JADBio optimizes the final model over a wealth of combinations of feature selection and classification algorithm combinations, along with their hyperparameter values, and estimates the predictive performance of the best-found model.

For prediction models to be applicable across clinical contexts, they must perform well when applied to data and populations of slightly different distributions than the one they were trained on [22]. However, this has not been tested in previous studies using ML to predict PTSD. Further, as already mentioned, few studies have aimed to predict PTSD symptoms over the long term (ie, several years) after traumatic events or military deployment. In this study, we aimed to address this by utilizing data routinely collected from 3 different cohorts deployed to Afghanistan with the Danish Defense that were followed for 2.5 and 6.5 years after returning from deployment. Specifically, in 3 experiments, we test if AutoML, as implemented in

JADBio, provides reliable performance estimates; if predictive models trained on one deployment cohort can be used to predict future screening-level PTSD in another deployment cohort; and how accurately we can predict screening-level PTSD 2.5 and 6.5 years after returning home using routinely collected questionnaire data. For the third aim, we report the model type that predicts the outcome best as well as the features selected as most predictive.

Methods

Population

The study population in this project included 3 different deployment cohorts that were similar in many regards in that all 3 cohorts were deployed to the same area in Afghanistan, each for a period of approximately 6 months between 2009 and 2013. However, the cohorts were different in some regards too, in that the mission purpose and level of threat were different across the deployments. All 3 deployment cohorts were part of the International Security Assistance Force (ISAF). In this study, cohort 1 (N=287 or N=261 depending on the timing of the outcome assessment, explained later) refers to ISAF7, who were deployed from February 2009 to August 2009; cohort 2 (N=352) refers to ISAF10, who were deployed from August 2010 to February 2011; and cohort 3 (N=232) refers to ISAF15, who were deployed from February 2013 to August 2013. Of note, cohort 1 is a subcohort of that used by Karstoft et al [17]; however, the data are different. Here, we used a new set of predictor data that was collected routinely for all 3 cohorts, but which was not part of the 2015 analysis. Descriptive statistics of the 3 cohorts can be seen in Table 1. In all 3 cohorts, most participants were male (>90%), and the mean age was 30.6-31.3 years. Prior deployment had occurred for 55.0%-63.6% of the cohorts, and 47.6%-59.2% of the cohorts had a military rank of private. Finally, the proportion with screening-level PTSD symptoms was 7.8%-10.0% 6 months after returning home, and 21.7%-27.3% were assessed as having PTSD 2.5 and 6.5 years after deployment.

Table 1. Descriptive statistics of the 3 cohorts.

Characteristics	Cohort 1[a] (N=261), n (%)	Cohort 2 (N=352), n (%)	Cohort 3 (N=232), n (%)
Age (years)[b]	30.6 (8.2)	31.3 (9.9)	31.1 (8.7)
Gender (female)	20 (7.7)	21 (6.0)	19 (8.2)
Previously deployed (yes)	162 (62.3)	193 (55.0)	147 (63.6)
Military rank (private)	154 (59.2)	193 (55.0)	110 (47.6)
Screening-level PTSD[c] at 6 months	22 (8.5)	35 (10.0)	18 (7.8)
Screening-level PTSD at the 2.5-year or 6.5-year follow-up	71 (27.3)	76 (21.7)	54 (23.4)

[a]Descriptive statistics for cohort 1 are based on the sample who provided outcome data at 6.5 years. Minor differences might be observed in the sample providing outcome data at 2.5 years.

[b]mean (SD).

[c]PTSD: posttraumatic stress disorder.

Data Material

Outcome

The predicted outcome was screening-level PTSD. For all 3 cohorts, this was assessed using the civilian version of the PTSD checklist (PCL-C) [23]. The PCL-C contains 17 items mirroring the symptoms of PTSD as defined in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition [24]. Our group has previously validated cutoff scores for the PCL-C in a military population and found that a score ≥44 identifies individuals with severe PTSD symptoms that indicates a likely PTSD diagnosis, while a score ≥30 can be used to identify individuals with moderate or screening-level PTSD [25]. For this study, we applied the cutoff score of 30 since the aim was not to identify individuals that most likely have a diagnosis but to screen for individuals that might be in need of help due to some elevation of symptomatology. For cohort 1, the outcome was assessed 2.5 years and 6.5 years after returning home. For cohort 2, the outcome was assessed 6.5 years after returning home only, while for cohort 3, it was assessed 2.5 years after returning home only. For the aim of this study, we combined the 3 cohorts in the following ways: cohorts 1 and 3 were combined (cohort 1&3_2.5, N=519) with the aim of predicting PTSD 2.5 years after deployment, while cohorts 1 and 2 were combined (cohort 1&2_6.5, N=613) with the aim of predicting PTSD 6.5 years after deployment.

Predictors

Predictors for the current project were retrieved from a database containing responses to the Psychological Reactions to International Missions (PRIM) questionnaire, which has been routinely distributed since 1998 to all Danish soldiers 6 months after return from an international deployment with the Danish Defense. The questionnaire contains 125 individual items covering deployment experiences (reported at 6 months after returning home), postdeployment reactions, and postdeployment support as well as 5 validated scales: PTSD symptoms [26], depression symptoms [27], perceived danger [28], witnessing of war atrocities during deployment [28], and postdeployment social support [29]. A list of all items in the PRIM questionnaire, translated into English (Multimedia Appendix 1), as well as their descriptive statistics including level of missingness (Multimedia Appendix 2 and Multimedia Appendix 3) can be seen in the supplementary material. Of note, the level of

missingness was <2% for all items except two, which had levels <3%.

Predictive Modeling With JADBio

For the predictive modeling in this project, we employed the AutoML program JADBio. JADBio has been employed in several other fields to produce novel scientific results (eg, nanomaterial property predictions [21], suicide prediction [30], speech classification [31], bank failure prediction [32], function protein prediction [33], and breast cancer prognosis and drug response prediction [34]). JADBio includes algorithms that are also appropriate for small-sample, high-dimensional biological data, hence the Bio part of the name, but can analyze any type of data that is in a 2-dimensional matrix format, as indicated by the examples provided. Internally, the system employs an artificial intelligence (AI) subsystem that encodes statistical knowledge to select the most appropriate algorithms for transformations, imputation of missing values, feature selection, and predictive modeling, as well as reasonable values for hyperparameters of these algorithms [21]. These selections are fed into what is called the Configuration Generator: Each configuration is a pipeline comprised of algorithms for transformations of features, imputation of missing values, feature selection, and modeling and corresponding hyperparameter values for each algorithm. Thus, a configuration accepts the data matrix and performs all steps necessary to generate a predictive model instance. Based on the choices of the AI system, the Configurator Generator searches in the space of possible configurations to identify one that is optimal, namely, the one that produces, on average, the best performing model instances. An important part of each configuration is the feature selection step, which is based on the statistical equivalent signatures algorithm [35]. The statistical equivalent signatures algorithm aims at identifying multiple feature subsets with the properties that they are of minimal size and optimally predictive for the target. Notice that multiple feature subsets may be equally predictive because of correlations among features. For example, a psychometric score computed on a few individual answers to a questionnaire may carry the same predictive informational content as some or all of the individual answers. For binary classification tasks, as in this work, JADBio employs standard statistical models (ridge logistic regression), non-statistical linear models (linear SVM), and non-linear models (decision trees, random forests, polynomial and Gaussian SVM). The final model is produced by applying the best performing configuration on all data. Thus, no samples are lost to estimation of performance.

To identify the winning configuration and estimate its predictive performance in an unbiased way, JADBio uses appropriate out-of-sample protocols (ie, protocols that hide some of the data from a configuration) and then evaluate the corresponding model on the held-out samples. Specifically, for small sample sizes, JADBio uses a stratified, *N* repeated, *K*-fold cross validation on each configuration. *K*-fold cross validation is a common estimation procedure. It partitions the available samples to *K* folds. It then feeds to a given configuration all folds but one and produces a model, which is then applied on the held-out fold to estimate performance (eg, AUC). The performance estimation of the models produced by the given configuration is the average over all folds. JADBio actually repeats the *K*-fold procedure *N* times and averages out the performance estimate. Each repetition randomly repartitions the data to different folds. The purpose of repeating cross validation is to reduce the variance of the estimate due to the specific partitioning to folds. “Stratification” is a specific variant of cross validation where each fold is constrained to have a class distribution (ie, percentage of PTSD cases vs. controls) similar to the class distribution in the original dataset. Stratification has been shown to reduce the variance of the estimation. The choice of the estimation protocol and the values of *N* and *K* are selected by the AI system based on the data characteristics (eg, sample size, imbalance of the classes) and the user preferences.

Notice that JADBio does not report the cross-validated estimate of the winning configuration: When one tries numerous configurations, the cross-validated estimate of the winner is overly optimistic [36]. For example, if the winning configuration has a cross-validated AUC of 0.9 out of 1000 other configurations, then the true expected AUC is likely to be <0.9. To remove the optimism, as well as compute confidence intervals of predictive performance, JADBio applies a method called Bootstrap Bias Corrected Cross-Validation. In general, the estimates returned by JADBio are conservative. The theory, algorithms, and empirical evidence for the estimation protocols are described in detail by Tsamardinos et al [36].

Modeling Procedure: Three Experiments to Test Performance Estimation, Model Transference, and Overall Prediction Accuracy

Experiment 1

First, we performed a computational experiment to ensure that JADBio’s predictive performance estimates are trustworthy. JADBio uses internally, theoretically, and empirically backed-up out-of-sample protocols to estimate performance of the final model, while adjusting for bias [36]. Nevertheless, it is still important to make sure the estimates of the system can be trusted in this particular type of data and problem. To this end, we initially combined the cohorts (cohorts 1 and 3 for 2.5-year PTSD prediction [cohort1&3_2.5] and cohorts 1 and 2 for 6.5-year PTSD prediction [cohort1&2_6.5]) and randomly split them into 5 subsets. Next, JADBio was used to train and evaluate models on four-fifths of the data and externally validate model performance on the remaining one-fifth of the data. We did this repeatedly to utilize all data subsets for training and validation, after which the performance achieved on the held-out fold was compared against the estimate returned by JADBio on the training folds.

Experiment 2

Second, we performed a computational experiment to establish transferability of the models (ie, test whether models trained on one cohort transfer [generalize] to another cohort). Specifically, we trained 2 models: one for PTSD status at 2.5 years after returning home and one for PTSD status at 6.5 years after returning home, both on data from cohort 1. We then tested their performance on cohort 3 and cohort 2, respectively.

Experiment 3

Third, we produced final models for each outcome and corresponding performance estimates from all available data. The reasoning behind the use of all available data was that, on average, the predictive performance of models increases with increased available sample size. Of course, this leaves no out-of-sample data to estimate the predictive performance; however, provided that experiment 1 was successful, the JADBio estimates can be trusted. In addition, provided that experiment 2 was successful, the model is likely to transfer to a new cohort and could potentially be employed in practice. More specifically, we again used the combined data sets (cohort1&3_2.5 and cohort1&2_6.5) to train models for each outcome. In addition, JADBio also performs feature selection during modeling. The selected features are the ones that enter the final models and provide psychological insight into the PTSD development.

Experiment 4 (Exploratory): Removal of Important Variables to Check Model Flexibility

While not part of the study aims, results from experiments 1-3 encouraged us to examine if the PTSD symptom level was the sole reason for the achieved prediction accuracy. Hence, we repeated experiment 3, but removed the total PTSD symptom score from the set of possible predictors. The purpose of this experiment was to test the robustness of JADBio in real-world screening situations, where some of the predictors can be missing. A desirable property of a screening method is to maintain predictive performance using available information.

Results

Experiment 1: Assessing the Quality of Out-of-Sample Performance Estimation

Results from the testing of JADBio performance estimates are displayed in Table 2. AUCs varied between 0.80 and 0.84 (mean 0.83) for the 2.5-year prediction and between 0.71 and 0.91 (mean 0.78) for the 6.5-year prediction. Importantly, the performances achieved on the validation sets are consistent with the JADBio performance estimates produced on the training sets; in fact, performance on the validation set was higher on average than the one estimated on the test sets. The results corroborate previous work [36], indicating that the estimation protocols within JADBio can be trusted on this data distribution. This implies that there is no need to reserve a separate hold-out set for estimating performance and lose samples to estimation.

Table 2. Areas under the receiver operating characteristic curves (AUCs) for the 5 training and test sets of the 2 cohorts.

Training-Testing set	2.5-year prediction		6.5-year prediction	
	Performance on the training set, AUC (95% CI)	Performance on test set, AUC	Performance on training set, AUC (95% CI)	Performance on test set, AUC
Training1-Testing1	0.77 (0.69-0.84)	0.82	0.77 (0.70-0.831)	0.73
Training2-Testing2	0.77 (0.68-0.84)	0.80	0.76 (0.69-0.83)	0.76
Training3-Testing3	0.73 (0.63-0.81)	0.84	0.74 (0.66-0.80)	0.91
Training4-Testing4	0.81 (0.73-0.88)	0.84	0.81 (0.74-0.87)	0.71
Training5-Testing5	0.78 (0.69-0.85)	0.82	0.74 (0.67-0.81)	0.79
Mean	0.77	0.83	0.76	0.78

Experiment 2: Assessing Model Transferability to Different Cohorts

For the 2.5-year threshold, the AUC of the best-found model was estimated at 0.76 (95% CI 0.67-0.84), while the AUC on the external validation cohort 3 was 0.79. For the 6.5-year threshold, the AUC of the best-found model was estimated at 0.70 (95% CI 0.60-0.80), while the AUC on the external validation cohort 2 was 0.81. The results provide evidence that models trained on one cohort transfer to a future cohort. Obviously, care needs to be applied with this statement for cohorts that are obtained far apart in time, on totally different populations, and for different military conflicts.

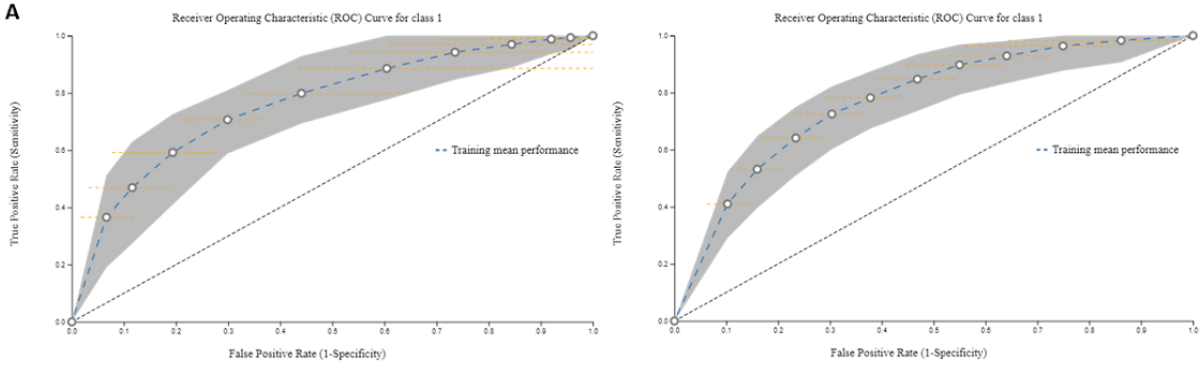
Experiment 3: Predictive Modeling for Potential Clinical Use

To produce the final predictive models for potential clinical use, we ran JADBio on cohort1&3 and cohort1&2 with PTSD status at 2.5 and 6.5 years, respectively, as the outcome. The user preferences were set to enforce feature selection and the analysis type to extensive, implying that a relatively large number of configurations would be explored. Overall, the analyses trained 450,200 and 417,800 models for the two outcomes, respectively, taking 32 and 40 minutes. All these models were trained using different configurations on different subsets of the data (cross validation) to estimate performance and produce a final optimal model. Detailed results from the prediction can be accessed in Multimedia Appendix 4.

For the 2.5-year prediction, the optimal model was a random forest classifier trained with 100 trees and a minimal number of observations per leaf of 5 (AUC=0.77, 95% CI 0.71-0.83). For the 6.5-year prediction, the optimal model was also a random forest classifier trained with 1000 trees and a minimal number of observations per leaf of 5 (AUC=0.78, 95% CI 0.73-0.83). ROC curves as well as sensitivity, specificity, positive predictive value, and negative predictive value for selected cutoffs can be seen in Figure 1A and Figure 1B, along with a confusion matrix for a suggested balanced cutoff (Figure 1C). The results indicate that 2.5-year, as well as 6.5-year, prognosis of PTSD is possible with applicable levels of predictive accuracy.

Figure 1. Results from the final prediction models (experiment 3): (A) receiver operating characteristic (ROC) curves for 2.5-year and 6.5-year predictions; (B) sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for selected cutoffs on the ROC curve; (C) confusion matrices for selected cutoffs on the ROC curve (marked in bold in Panel B).

A



B

2.5-year prediction

Threshold	Sensitivity	Specificity	PPV	NPV
0.44	0.37	0.93	0.63	0.82
0.36	0.47	0.88	0.55	0.84
0.28	0.59	0.81	0.50	0.86
0.22	**0.71**	**0.70**	**0.43**	**0.88**
0.17	0.80	0.56	0.36	0.90
0.13	0.89	0.40	0.32	0.92

6.5-year prediction

Threshold	Sensitivity	Specificity	PPV	NPV
0.39	0.41	0.90	0.56	0.83
0.33	0.53	0.84	0.51	0.85
0.27	0.64	0.77	0.47	0.87
0.23	**0.73**	**0.70**	**0.43**	**0.89**
0.20	0.78	0.62	0.39	0.90
0.17	0.85	0.53	0.36	0.92

C

Observed Class	Predicted Class: No	Predicted Class: Yes
No	0.53 [0.46, 0.60]	0.23 [0.16, 0.28]
Yes	0.07 [0.04, 0.10]	0.17 [0.13, 0.21]

Observed Class	Predicted Class: No	Predicted Class: Yes
No	0.53 [0.47, 0.59]	0.23 [0.17, 0.28]
Yes	0.07 [0.04, 0.10]	0.17 [0.14, 0.21]

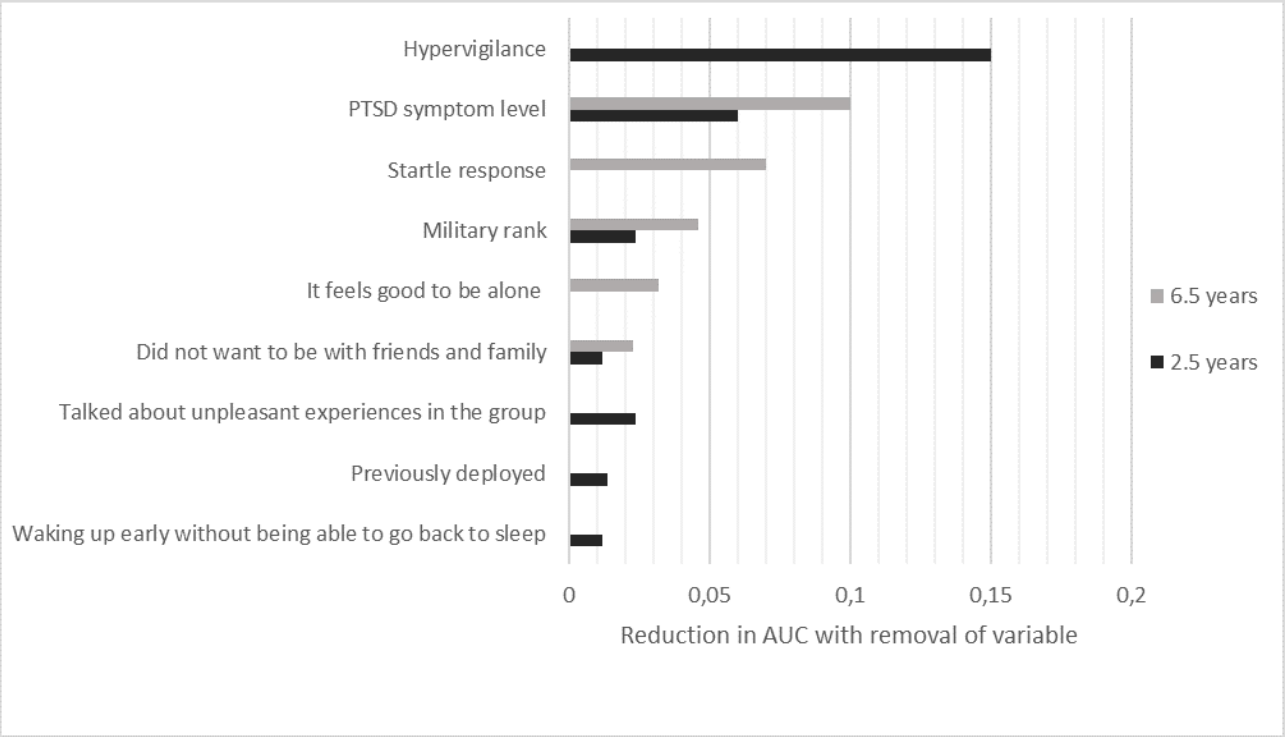
Experiment 3: Comparison Between Linear and Non-Linear Models

While random forests were the best overall performing models for both outcomes, JADBio also reported the best performing models out of those that are humanly interpretable. These were generalized linear models (ridge logistic regression) and decision trees. The best interpretable models for each outcome were both ridge logistic regression models. Their predictive performance was estimated to be indistinguishable from the random forests. The results show that, in these predictive tasks, non-linear models do not significantly improve predictive performance.

Experiment 3: Feature Selection

For the 2.5-year prediction, 4 similar, equally predictive feature sets were selected, each consisting of 14 features and, in total, including 16 different features across the 4 feature sets. For the 6.5-year prediction, a single, optimal feature subset was discovered consisting of 9 features (see Multimedia Appendix 5 for a total list of selected features). Variable importance is depicted in Figure 2, which includes all selected features of both models that lead to an AUC reduction of at least 0.01 if removed from the model. For both cohorts, total level of PTSD symptoms, military rank, and little desire to be with friends and family led to reductions in AUC if removed. Further, in both models, having a hyperarousal symptom had relatively large importance; for the 2.5-year prediction, hypervigilance showed the greatest importance, whereas for the 6.5-year prediction, a startle response led to a relatively large reduction in AUC if removed.

Figure 2. Cumulative variable importance for the two final prediction models. The x axis depicts the reduction in the area under the receiver operating characteristic curve (AUC) if the particular variable is removed from the model. The figure includes all selected variables resulting in an AUC reduction of at least 0.01 if removed. PTSD: posttraumatic stress disorder.



Experiment 4 (Exploratory): Removal of Important Variables to Test Model Flexibility

For the 2.5-year as well as the 6.5-year prediction, the prediction accuracy of the models when removing the total level of PTSD symptoms remained the same (2.5-year prediction, AUC=0.77, 95% CI 0.71-0.82; 6.5-year prediction, AUC=0.78, 95% CI 0.72-0.84). The results indicate that, even when removing a central feature, predictive accuracy can be maintained.

Discussion

For predictive models of PTSD symptomatology following military deployment to be useful in practical settings, several things are important: Applied models are not overly optimistic, predictive models built on current deployment cohorts can be transferred to future cohorts, and sufficiently high predictive accuracy can be reached for long-term PTSD outcomes (ie, several years following deployment). Testing and optimizing models manually are time-consuming; therefore, this study tested the applicability of AutoML as a means of enhancing model selection and parameter optimization. Hence, in the current study, we aimed to evaluate if this was achievable by combining data from 3 different cohorts deployed to Afghanistan with the Danish Defense between 2009 and 2013 to build predictive models using automated ML methods. Overall, we found that our applied AutoML software produced reliable estimates, the identified predictive models transferred well, and acceptable predictive accuracies were reached for prediction of screening-level PTSD 2.5 years after deployment (AUC=0.77, 95% CI 0.71-0.83) and 6.5 years after deployment (AUC=0.78, 95% CI 0.73-0.83). Further, we found that linear and nonlinear models performed equally well, and that, even with removal of one of the most central features, namely the total level of PTSD symptoms, screening-level PTSD at 2.5 years and 6.5 years could be predicted.

The use of an AutoML program such as JADBio warrants some discussion. A major advantage of using such a program is that it allows us to test multiple combinations of algorithms and their hyperparameter values within a reasonable time frame without need of extensive computer power. One drawback is that we have to rely on the settings of JADBio in performance evaluation. Here, one might worry that JADBio could be overestimating performance of identified models, for example by insufficient correction for the multitude of tested models [37]. To test if this was the case, in experiment 1, we performed a 5-fold cross validation of our two combined data sets where one-fifth of the data were repeatedly held out for external validation. Reassuringly, for all 10 validations, we found that the external validations of the test set revealed similar prediction accuracy as for the training set — all except three within the training set Cis, one slightly below, two slightly above. Hence, it seems reasonable to conclude that performance evaluation in JADBio is not overly optimistic, at least for the size and type of data examined here. This assured us that we can apply JADBio on all samples available for a given task without having to withhold a separate validation (hold-out) set and lose samples to estimation.

Having established that, model transferability was the next important prerequisite for the successful implementation of ML-based screening of deployment cohorts based on routinely collected data. Our results suggest that predictive models built on one cohort can indeed transfer to other cohorts. When predicting screening-level PTSD 2.5 years after deployment, our results showed that the model trained and tested on cohort

1 performed with similar accuracy in cohort 3 (cohort 3 AUC=0.79; cohort 1 AUC=0.76, 95% CI 0.67-0.84). When predicting screening-level PTSD 6.5 years after deployment, our results actually suggested that the model trained and tested on cohort 1 performed better on cohort 2 (cohort 2 AUC=0.81; cohort 1 AUC=0.70, 95% CI 0.60-0.80). Hence, for the deployment cohorts included in this study, it seems safe to say that models trained and tested on one cohort can transfer to another cohort. Importantly, while all cohorts included in this study deployed to Afghanistan, they did so at different times, with cohort 1 deploying in 2009 and cohort 3 in 2013. Conditions, tasks, threat levels, and deployment environments were similar between cohorts 1 and 2 [38] while substantially different between cohorts 1 and 3 [39], suggesting that even when deployment characteristics are not the same, predictive models can be transferred between cohorts deployed in similar missions. This is important given a wish to apply predictive models identified on existing data to future deployment cohorts.

This is one of the first few studies that tests how accurately PTSD can be predicted several years following deployment, more accurately, 2.5 and 6.5 years after returning home. From the literature, we know that symptoms of PTSD might develop with some delay following trauma, especially when the trauma occurs in an occupational context such as during military deployment [40,41]. Hence, predictive models trained to identify people who develop symptoms only during the first months following deployment might miss a great deal of those who go on to experience PTSD symptomatology. Our models predict screening-level PTSD at 2.5 years and 6.5 years with similar, acceptable accuracy (AUCs=0.77 and 0.78, respectively). Further, based on the model values of sensitivity, specificity, positive predictive value, and negative predictive value (Table 1), our findings show that we can achieve a reasonable balance for screening purposes. For example, with a sensitivity of 0.73 (for 6.5-year prediction), 89% of those who screen negative will indeed be noncases, while 43% of those who screen positive will indeed be cases. Optimally, our models would have higher overall accuracy; however, we utilized routinely collected data that were not collected with prediction in mind, and we were interested in testing how accurately prediction models trained on these data could predict screening-level PTSD. While the prediction accuracy is acceptable, it is far from perfect, and future endeavors should preferably include features that might increase accuracy.

From the total number of features, relatively small features sets were selected for both predictive models (14 features for 2.5-year prediction, 9 features for 6.5-year prediction). Some overlap in selected features was seen, with 3 features showing high cumulative importance in both cohorts: military rank, diminished interest in being with friends and family, and total level of PTSD symptoms 6 months after returning home. Neither of these are surprising; Lower rank has consistently been identified as a risk factor for PTSD following military trauma [42], low levels of perceived social support following trauma exposure is a known risk factor for PTSD [43], and early post-trauma levels of PTSD symptoms has also been found to predict PTSD later on [6].

Further, we found that a hyperarousal symptom is important in both cohorts: For the 2.5-year prediction, hypervigilance was the single most important feature, leading to a 0.15 reduction in AUC if removed, while for the 6.5-year prediction, startle response was among the most important features. While it is somewhat surprising that individual hyperarousal symptoms were selected as predictive features over and above the total level of PTSD severity, hypervigilance and startle response have been found to be central symptoms in PTSD in previous research [44,45]. This finding illustrates one of the benefits of using ML approaches for predictive modeling: In linear approaches, overlapping constructs would likely go undetected as significant, individual predictors of the outcome. The Bayesian Network approach for feature selection implemented in JADBio clearly allows for detection of such related predictors that are nonetheless individually related to the outcome [35].

An analyst facing predictive modeling tasks does not know whether interpretable, standard statistical linear models suffice or more complex, nonlinear, ML-based models are necessary to achieve optimal predictive performance. JADBio automatically tries both types of models and allows an analyst to compare them on equal grounds. An analyst can thus gauge whether the use of complex, nonlinear models is justified by the increase in predictive performance achieved. In our analyses, it seems that the linear models performed equally well for any practical purpose.

Also, we found that even when removing one of the most important predictive features, the total level of PTSD symptoms at 6 months, the prediction accuracy did not decrease. Hence, even when total symptom level is not available at 6 months, screening-level PTSD at 2.5 and 6.5 years can be predicted. This implies that even with a limited number of individual symptoms, personal characteristics, and demographic variables available, we will be able to identify those who have the highest risk of developing screening-level PTSD.

Our study has some limitations that should be noted. First, while we included 3 different cohorts, they are similar in that they all deployed to Afghanistan. Hence, to test if models transfer also when, for example, the deployment country differs, we will need to include cohorts who deployed to other conflict zones. However, as already argued, the cohorts differed in important ways. Second, the response rate to the questionnaire was approximately 65% across cohorts. We know from earlier analyses based on these cohorts that more individuals among the nonresponders may have more mental health problems [3]. While this introduces a risk of bias, this is also the reality if future screening efforts are to be based on this approach: For all Danish deployment cohorts since 1998, the response rate varies around 65%, so screening will be limited to those who responded to the questionnaire. Third, since the PRIM questionnaire is already being used for screening, individuals might have been offered treatment as a result of their responses to PRIM, which is another source of potential bias.

Despite these limitations, our study illustrates how screening of future deployment cohorts can be based on ML-based predictive models based solely on routinely collected questionnaire data. Importantly, we have demonstrated that

models developed on routinely collected data on one cohort can be successfully transferred to predict screening-level PTSD in another cohort deployed to similar missions and that satisfactory predictive accuracy can be reached such that the model can be used as an actual decision support tool. In future efforts, we suggest that the models are further validated in cohorts deployed to other missions. For cohort 3 of this study, a follow-up collection of post-deployment data including measurement of PTSD symptoms is being conducted in the spring and summer of 2020. We predict that, based on a model trained on our routinely collected data at 6 months after homecoming for cohorts 1 and 2, we will be able to classify screening-level PTSD 6.5 years after returning home in cohort 3 with an AUC between 0.73 and 0.83. We intend to preregister and publish the results of this endeavor.

Acknowledgments

The authors would like to thank all deployed personnel from the three cohorts for their participation in this study. Further, the authors thank Dr. Anni B. S. Nielsen for her role in data collection and preparation.

Conflicts of Interest

IT is the founder and CEO of Gnosis Data Analysis PC.

Multimedia Appendix 1
Full list of predictors.
[DOCX File , 20 KB - medinform_v8i7e17119_app1.docx]

Multimedia Appendix 2
Descriptive characteristics of all variables in cohorts 1 and 2.
[TXT File , 58 KB - medinform_v8i7e17119_app2.txt]

Multimedia Appendix 3
Descriptive characteristics of all variables in cohorts 1 and 3.
[TXT File , 59 KB - medinform_v8i7e17119_app3.txt]

Multimedia Appendix 4
Detailed results and visualizations from JADBio.
[DOCX File , 12 KB - medinform_v8i7e17119_app4.docx]

Multimedia Appendix 5
Selected features for 2.5-year and 6.5-year predictions.
[DOCX File , 14 KB - medinform_v8i7e17119_app5.docx]

References

1. Yehuda R, Hoge CW, McFarlane AC, Vermetten E, Lanius RA, Nievergelt CM, et al. Post-traumatic stress disorder. Nat Rev Dis Primers 2015 Oct 8;1(1). [doi: 10.1038/nrdp.2015.57]
2. Hines LA, Sundin J, Rona RJ, Wessely S, Fear NT. Posttraumatic stress disorder post Iraq and Afghanistan: prevalence among military subgroups. Can J Psychiatry 2014 Sep;59(9):468-479 [FREE Full text] [doi: 10.1177/070674371405900903] [Medline: 25569079]
3. Madsen T, Andersen SB, Karstoft K. Are Posttraumatic Stress Symptoms Related to Mental Health Service Use? J. Clin. Psychiatry 2016 Aug 16;77(10):e1226-e1232. [doi: 10.4088/jcp.15m10088]
4. Lee DJ, Warner CH, Hoge CW. Advances and controversies in military posttraumatic stress disorder screening. Curr Psychiatry Rep 2014 Sep;16(9):467. [doi: 10.1007/s11920-014-0467-7] [Medline: 25023512]
5. Rona RJ, Burdett H, Khondoker M, Chesnokov M, Green K, Pernet D, et al. Post-deployment screening for mental disorders and tailored advice about help-seeking in the UK military: a cluster randomised controlled trial. The Lancet 2017 Apr;389(10077):1410-1423. [doi: 10.1016/s0140-6736(16)32398-4]
6. Shalev AY, Gevonden M, Ratanatharathorn A, Laska E, van der Mei WF, Qi W, International Consortium to Predict PTSD. Estimating the risk of PTSD in recent trauma survivors: results of the International Consortium to Predict PTSD (ICPP). World Psychiatry 2019 Feb 02;18(1):77-87 [FREE Full text] [doi: 10.1002/wps.20608] [Medline: 30600620]
7. Breiman L. Statistical Modeling: The Two Cultures. Statistical Science ? 2001;16(3):215.

8. Schultebraucks K, Galatzer-Levy IR. Machine Learning for Prediction of Posttraumatic Stress and Resilience Following Trauma: An Overview of Basic Concepts and Recent Advances. J Trauma Stress 2019 Apr 20;32(2):215-225. [doi: 10.1002/jts.22384] [Medline: 30892723]
9. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. Mol Psychiatry 2017 Jan;22(1):37-43. [doi: 10.1038/mp.2016.201] [Medline: 27843153]
10. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med 2016 Sep;46(12):2455-2465 [FREE Full text] [doi: 10.1017/S0033291716001367] [Medline: 27406289]
11. Bishop C. Pattern recognition and machine learning. New York: Springer; 2009:978.
12. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition. New York: Springer-Verlag; 2009.
13. Saxe GN, Ma S, Ren J, Aliferis C. Machine learning methods to predict child posttraumatic stress: a proof of concept study. BMC Psychiatry 2017 Jul 10;17(1). [doi: 10.1186/s12888-017-1384-1]
14. Rosellini AJ, Dussaillant F, Zubizarreta JR, Kessler RC, Rose S. Predicting posttraumatic stress disorder following a natural disaster. J Psychiatr Res 2018 Jan;96:15-22 [FREE Full text] [doi: 10.1016/j.jpsychires.2017.09.010] [Medline: 28950110]
15. Galatzer-Levy IR, Karstoft K, Statnikov A, Shalev AY. Quantitative forecasting of PTSD from early trauma responses: a Machine Learning application. J Psychiatr Res 2014 Dec;59:68-76 [FREE Full text] [doi: 10.1016/j.jpsychires.2014.08.017] [Medline: 25260752]
16. Karstoft K, Galatzer-Levy IR, Statnikov A, Li Z, Shalev AY, members of Jerusalem Trauma Outreach Prevention Study (J-TOPS) group. Bridging a translational gap: using machine learning to improve the prediction of PTSD. BMC Psychiatry 2015 Mar 16;15(1):30 [FREE Full text] [doi: 10.1186/s12888-015-0399-8] [Medline: 25886446]
17. Karstoft K, Statnikov A, Andersen SB, Madsen T, Galatzer-Levy IR. Early identification of posttraumatic stress following military deployment: Application of machine learning methods to a prospective study of Danish soldiers. J Affect Disord 2015 Sep 15;184:170-175. [doi: 10.1016/j.jad.2015.05.057] [Medline: 26093830]
18. Kessler RC, Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, Army STARRS Collaborators. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study To Assess Risk and rEsilience in Servicemembers (Army STARRS). JAMA Psychiatry 2015 Jan;72(1):49-57 [FREE Full text] [doi: 10.1001/jamapsychiatry.2014.1754] [Medline: 25390793]
19. Rosellini AJ, Stein MB, Benedek DM, Bliese PD, Chiu WT, Hwang I, et al. Predeployment predictors of psychiatric disorder-symptoms and interpersonal violence during combat deployment. Depress Anxiety 2018 Nov 13;35(11):1073-1080 [FREE Full text] [doi: 10.1002/da.22807] [Medline: 30102442]
20. Guyon I, Bennett K, Cawley G, Escalante H, Escalera S, Tin KH, et al. Design of the 2015 ChaLearn AutoML challenge. In: Design of the 2015 ChaLearn AutoML challenge. 2015 International Joint Conference on Neural Networks (IJCNN) Internet Killarney, Ireland: IEEE; 2015 Presented at: International Joint Conference on Neural Networks; 2015-11-07; Killarney, Ireland. [doi: 10.1109/ijcnn.2015.7280767]
21. Borboudakis G, Stergiannakos T, Frysali M, Klontzas E, Tsamardinos I, Froudakis GE. Author Correction: Chemically intuited, large-scale screening of MOFs by machine learning techniques. npj Comput Mater 2017 Nov 8;3(1). [doi: 10.1038/s41524-017-0051-x]
22. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Mach Learn 2009 Oct 23;79(1-2):151-175. [doi: 10.1007/s10994-009-5152-4]
23. Weathers F, Litz B, Herman D, Huska J, Keane T. The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility. 1993 Presented at: International Society of Traumatic Stress Studies; 1993; San Antonio, Texas, US.
24. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: Fourth edition. Washington DC: American Psychiatric Association; 1994.
25. Karstoft K, Andersen SB, Bertelsen M, Madsen T. Diagnostic accuracy of the posttraumatic stress disorder checklist-civilian version in a representative military sample. Psychol Assess 2014 Mar;26(1):321-325. [doi: 10.1037/a0034889] [Medline: 24188155]
26. Karstoft K, Andersen SB, Nielsen ABS. Assessing PTSD in the military: Validation of a scale distributed to Danish soldiers after deployment since 1998. Scand J Psychol 2017 Apr 18;58(3):260-268. [doi: 10.1111/sjop.12360]
27. Karstoft K, Nielsen ABS, Nielsen T. Assessment of depression in veterans across missions: a validity study using Rasch measurement models. Eur J Psychotraumatol 2017 May 22;8(1):1326798 [FREE Full text] [doi: 10.1080/20008198.2017.1326798] [Medline: 28649301]
28. Karstoft K, Nielsen T, Nielsen ABS. Perceived danger during deployment: a Rasch validation of an instrument assessing perceived combat exposure and the witnessing of combat consequences in a war zone. Eur J Psychotraumatol 2018 Jul 09;9(1):1487224 [FREE Full text] [doi: 10.1080/20008198.2018.1487224] [Medline: 30013725]
29. Karstoft K, Nielsen T, Nielsen ABS. Measuring Social Support among Soldiers with the Experienced Post-Deployment Social Support Scale (EPSSS): A Rasch-Based Construct Validity Study. Behav Med 2019 Oct 16:1-9. [doi: 10.1080/08964289.2019.1676192] [Medline: 31617826]
30. Adamou M, Antoniou G, Greasidou E, Lagani V, Charonyktakis P, Tsamardinos I, et al. Toward Automatic Risk Assessment to Support Suicide Prevention. Crisis 2019 Jul;40(4):249-256. [doi: 10.1027/0227-5910/a000561] [Medline: 30474411]

31. Simantiraki O, Charonyktakis P, Pampouchidou A, Tsiknakis M, Cooke M. Glottal source features for automatic speech-based depression assessment. 2017 Presented at: Proceedings of the 18th Conference of the International Speech Communication Association INTERSPEECH; 2017; Stockholm, Sweden p. 2700. [doi: 10.21437/interspeech.2017-1251]
32. Agrapetidou A, Charonyktakis P, Gogas P, Papadimitriou T, Tsamardinos I. An AutoML application to forecasting bank failures. Applied Economics Letters 2020 Feb 03:1-5. [doi: 10.1080/13504851.2020.1725230]
33. Orfanoudaki G, Markaki M, Chatzi K, Tsamardinos I, Economou A. MatureP: prediction of secreted proteins with exclusive information from their mature regions. Sci Rep 2017 Jun 12;7(1):3263 [FREE Full text] [doi: 10.1038/s41598-017-03557-4] [Medline: 28607462]
34. Panagopoulou M, Karaglani M, Balgkouranidou I, Biziota E, Koukaki T, Karamitrousis E, et al. Circulating cell-free DNA in breast cancer: size profiling, levels, and methylation patterns lead to prognostic and predictive classifiers. Oncogene 2019 May;38(18):3387-3401. [doi: 10.1038/s41388-018-0660-y] [Medline: 30643192]
35. Tsagris M, Tsamardinos I. Feature selection with the R package. F1000Res 2018;7:1505 [FREE Full text] [Medline: 31656581]
36. Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach Learn 2018 May 9;107(12):1895-1922. [doi: 10.1007/s10994-018-5714-4]
37. Jensen D, Cohen P, Cohen P. Multiple Comparisons in Induction Algorithms. Mach Learn 2000;38(3):338. [doi: 10.1023/A:1007631014630]
38. Nielsen A, Andersen S, Karstoft KI. ISAF10 - 6,5 år efter hjemkomst. Ringsted: Veterancentret; 2019.
39. Karstoft K, Nielsen A, Andersen S. ISAF7 - 6,5 år efter hjemkomst. In: Veterancentret. Ringsted: Veterancentret; 2017.
40. Andrews B, Brewin CR, Philpott R, Stewart L. Delayed-Onset Posttraumatic Stress Disorder: A Systematic Review of the Evidence. AJP 2007 Sep;164(9):1319-1326. [doi: 10.1176/appi.ajp.2007.06091491]
41. Utzon-Frank N, Breinegaard N, Bertelsen M, Borritz M, Eller NH, Nordentoft M, et al. Occurrence of delayed-onset post-traumatic stress disorder: a systematic review and meta-analysis of prospective studies. Scand J Work Environ Health 2014 Mar 06;40(3):215-229. [doi: 10.5271/sjweh.3420]
42. Jones M, Sundin J, Goodwin L, Hull L, Fear NT, Wessely S, et al. What explains post-traumatic stress disorder (PTSD) in UK service personnel: deployment or something else? Psychol. Med 2012 Nov 13;43(8):1703-1712. [doi: 10.1017/s0033291712002619]
43. Brewin CR, Andrews B, Valentine JD. Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. J Consult Clin Psychol 2000 Oct;68(5):748-766. [Medline: 11068961]
44. Greene T, Gelkopf M, Epskamp S, Fried E. Dynamic networks of PTSD symptoms during conflict. Psychol. Med 2018 Feb 28;48(14):2409-2417. [doi: 10.1017/s0033291718000351]
45. Schell TL, Marshall GN, Jaycox LH. All symptoms are not created equal: the prominent role of hyperarousal in the natural course of posttraumatic psychological distress. J Abnorm Psychol 2004 May;113(2):189-197. [doi: 10.1037/0021-843X.113.2.189] [Medline: 15122939]

Abbreviations

AI: artificial intelligence.
AUC: area under the receiver operating characteristic curve.
AutoML: automated machine learning.
ISAF: International Security Assistance Force.
JADBio: Just Add Data Bio.
ML: machine learning.
PCL-C: civilian version of the PTSD checklist.
PRIM: Psychological Reactions to International Missions.
PTSD: posttraumatic stress disorder.
ROC: receiver operating characteristic.
SVM: support vector machine.

Edited by C Lovis; submitted 20.11.19; peer-reviewed by K Schultebraucks, M van Zuiden; comments to author 19.01.20; revised version received 30.03.20; accepted 16.04.20; published 22.07.20.

Please cite as:
Karstoft KI, Tsamardinos I, Eskelund K, Andersen SB, Nissen LR
Applicability of an Automated Model and Parameter Selection in the Prediction of Screening-Level PTSD in Danish Soldiers Following Deployment: Development Study of Transferable Predictive Models Using Automated Machine Learning
JMIR Med Inform 2020;8(7):e17119
URL: http://medinform.jmir.org/2020/7/e17119/
doi:10.2196/17119
PMID:32706722

Original Paper

Predicting the Development of Type 2 Diabetes in a Large Australian Cohort Using Machine-Learning Techniques: Longitudinal Survey Study

Lei Zhang[1], PhD; Xianwen Shang[2], PhD; Subhashaan Sreedharan[2], MD; Xixi Yan[2], PhD; Jianbin Liu[2], MD; Stuart Keel[2], PhD; Jinrong Wu[2], MA; Wei Peng[3], PhD; Mingguang He[2], PhD

[1]China-Australia Joint Research Center for Infectious Diseases, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, Shaanxi, China

[2]Centre for Eye Research Australia; Ophthalmology, Department of Surgery, The University of Melbourne, Melbourne, Australia

[3]Research Centre for Data Analytics and Cognition, La Trobe University, Melbourne, Australia

Corresponding Author:
Lei Zhang, PhD
China-Australia Joint Research Center for Infectious Diseases
School of Public Health
Xi'an Jiaotong University Health Science Center
76 Yanta West Road, Yanta District
Xi'an, Shaanxi, 710061
China
Phone: 86 15910593477
Email: Lei.Zhang1@monash.edu

Abstract

Background: Previous conventional models for the prediction of diabetes could be updated by incorporating the increasing amount of health data available and new risk prediction methodology.

Objective: We aimed to develop a substantially improved diabetes risk prediction model using sophisticated machine-learning algorithms based on a large retrospective population cohort of over 230,000 people who were enrolled in the study during 2006-2017.

Methods: We collected demographic, medical, behavioral, and incidence data for type 2 diabetes mellitus (T2DM) in over 236,684 diabetes-free participants recruited from the 45 and Up Study. We predicted and compared the risk of diabetes onset in these participants at 3, 5, 7, and 10 years based on three machine-learning approaches and the conventional regression model.

Results: Overall, 6.05% (14,313/236,684) of the participants developed T2DM during an average 8.8-year follow-up period. The 10-year diabetes incidence in men was 8.30% (8.08%-8.49%), which was significantly higher (odds ratio 1.37, 95% CI 1.32-1.41) than that in women at 6.20% (6.00%-6.40%). The incidence of T2DM was doubled in individuals with obesity (men: 17.78% [17.05%-18.43%]; women: 14.59% [13.99%-15.17%]) compared with that of nonobese individuals. The gradient boosting machine model showed the best performance among the four models (area under the curve of 79% in 3-year prediction and 75% in 10-year prediction). All machine-learning models predicted BMI as the most significant factor contributing to diabetes onset, which explained 12%-50% of the variance in the prediction of diabetes. The model predicted that if BMI in obese and overweight participants could be hypothetically reduced to a healthy range, the 10-year probability of diabetes onset would be significantly reduced from 8.3% to 2.8% (P<.001).

Conclusions: A one-time self-reported survey can accurately predict the risk of diabetes using a machine-learning approach. Achieving a healthy BMI can significantly reduce the risk of developing T2DM.

(JMIR Med Inform 2020;8(7):e16850) doi:10.2196/16850

KEYWORDS

diabetes; machine learning; risk prediction; cohort study

Introduction

Diabetes and its complications are major causes of premature mortality globally. It is estimated that 451 million people worldwide had diabetes in 2017, and this figure is projected to rise by 35% to 693 million by 2045 [1]. In addition to the disease burden of diabetes, the annual global economic cost associated with diabetes is currently estimated to be US $1.3 trillion [2].

Predicting the risk of diabetes in adults has been a primary focus in many health care systems internationally. In the last 20 years, numerous diabetes risk prediction tools have been developed with variable success [3-12]. Among these, four were published by national government agencies (United States [10], Australia [11], United Kingdom [9], and Canada [8]) and are freely accessible online. The vast majority of these tools collect information on individual demographical characteristics, medical history, family history, anthropometric measurements, and biomarkers, and produce a "risk score" based on regression models. However, these conventional models share some major shortcomings. First, all of these tools include blood glucose level as a predictor, which leads to spurious inflated prediction accuracy because the glucose level per se defines diabetes. Prediction based on a predicting factor that defines outcomes will inevitably achieve high accuracy. Second, these tools have been developed based on relatively small sample sizes (typically 5200-6400 individuals) and include participants recruited from only select communities. Third, the datasets utilized are outdated and therefore represent a potential source of bias. For example, the American Diabetes Association Questionnaire is based on the National Health and Nutrition Examination conducted during 1999-2004 [10] and the Australian Type 2 Diabetes Risk Assessment Tool is based on the 1999-2000 AusDiab-Australian Diabetes, Obesity and Lifestyle study [11]. Fourth, all of these tools employed a conventional regression model for risk prediction.

Therefore, these models could be updated by incorporating the increasing amount of health data available and new risk prediction methodology available to date. Interestingly, the 2014 EPIC-InterACT study reviewed and validated 12 conventional prediction models based on a case-cohort sample of 27,779 European individuals [12]. The results suggested that these models can identify individuals at high risk of developing type 2 diabetes mellitus (T2DM), but the performance of the models varied substantially with country, age, sex, and body weight. More recently, the QDiabetes study led by Hippisley-Cox et al [13] overcame many of these shortcomings. Based on a large population dataset of 11.5 million individuals, this model provides a 10-year risk prediction for diabetes with the option to include or exclude fasting blood glucose and glycated hemoglobin as predictors. Despite this progress, the study employed a conventional Cox proportional hazards model, which suffers from some major limitations associated with its assumptions in which the predictors are assumed to have time-independent and linear impacts on the hazard.

Machine learning is an emerging and widely accepted approach for risk prediction [14]. Various machine-learning algorithms have been proposed, ranging from conventional to more advanced ensemble machine-learning approaches [15]. However, a shared common trait in most models is reliance on the presence of biomarkers. For instance, the blood glucose level is a biomarker that is commonly adopted in several machine-learning models with an estimated area under the receiver operating characteristic curve (AUC) value in the 70%-80% range [16-18]. Combining the information of both blood glucose levels and other biological parameters has been shown to improve the machine-learning accuracy [19], but the collection of biomarkers requires invasive blood sampling and is limited to clinical settings. Therefore, development of an accurate prediction tool that solely depends on self-reported information offers great potential for wider application in resource-limited settings to combat the growing global diabetes epidemic.

We argue that a new risk prediction tool is needed to address the shortcomings of current tools. Toward this end, in this study, we present a machine learning–based diabetes risk prediction tool using only self-reported information. This model was based on data from a large cohort of more than 230,000 residents in New South Wales (NSW), Australia collected during the period of 2006-2017. More specifically, the tool aims to address two questions. First, can the risk of diabetes be predicted in both the short and long term (3-10 years) based on a one-time self-reported survey without any biomarkers? Second, can the effects of modifiable risk factors for diabetes onset be assessed with such a tool?

Methods

The 45 and Up Study

The Sax Institute's 45 and Up Study is the largest prospective cohort study conducted in Australia [20]. This study enrolled 266,896 residents aged 45 years and older from NSW, Australia between 2006 and 2009, representing around 11% of the NSW population in this age group [20]. The study methodology has been described in detail elsewhere [20]. Eligible participants aged 45 and over and residents of NSW were randomly sampled from the Medicare Australia enrolment database, and received an invitation by mail including a study questionnaire and a written informed consent form. All participants provided consent for linkage of their information to routine health databases. The baseline questionnaire captured information on a broad range of socioeconomic, health, and lifestyle factors. To track medical procedures and medications received by the participants, the 45 and Up Study data was linked to the Medicare Benefits Schedule and Pharmaceutical Benefits Scheme claims from 2004 to 2016 using a unique identifier provided by the Department of Human Services. The Medicare Benefits Schedule code is a unique identifying code for medical procedures, whereas the Pharmaceutical Benefits Scheme is the identifying code for medications prescribed by clinicians.

Ethical Considerations

Ethics approval of the 45 and Up Study was obtained from the University of New South Wales Human Research Ethics Committee. Approval to use data from the 45 and Up Study for the current study was received from the Royal Victorian Eye and Ear Hospital Human Research Ethics Committee.

Inclusion and Exclusion Criteria

We excluded participants with established diabetes at baseline, defined as those who: (1) provided a positive response to question no. 24 “Has a doctor EVER told you that you have diabetes?”; (2) used diabetes medications based on the Pharmaceutical Benefits Scheme database before the baseline survey [21]; or (3) had gestational diabetes, defined as a diagnosis of diabetes earlier than the last childbirth, but without diabetes medication use subsequently. We also excluded participants who had incomplete physical activity data, and those who reported an age of diabetes diagnosis older than the age at the baseline survey. Among the 266,896 participants from the 45 and Up Study, we included a total of 236,584 residents in this study (Multimedia Appendix 1).

Key Outcome and Predicting Variables

The primary outcome of the study was the first occurrence of prescription for any kind of medication for T2DM (including oral hypoglycemic agents and insulin). Prescription of a diabetes medication was defined as the corresponding Pharmaceutical Benefits Scheme codes detailed in Multimedia Appendix 2. As all participants were aged >45 years, we assumed that all cases of new diabetes medication use were for T2DM rather than type 1 diabetes mellitus. We intended to project the risk of diabetes with a one-time self-reported survey at baseline (Multimedia Appendix 3), which included no biomarkers such as blood glucose levels. The four categories of a total of 39 predicting variables included: demographic characteristics, medical and family history, lifestyle indicators, and dietary indicators. We acknowledge that our definition of T2DM may likely overlook cases of gestational diabetes.

Conventional Regression Model

We employed a conventional logistic regression model to investigate the incidence of diabetes and its association with the predicting variables. We investigated the risk of diabetes and its associated factors for a duration of 3, 5, 7, and 10 years after baseline using four separate models. For each of these models, only participants who were part of the respective follow-up duration were included. We used the conventional regression model as the benchmark model as it is well established to be the standard method for investigating associations between a binary outcome and potential relevant factors.

Machine-Learning Models

For comparison with the regression model, we applied three commonly used machine-learning models, which included a random forest, multilayer feedforward artificial neural network implementing a deep-learning approach, and a gradient boosting machine approach. These three models represent the mainstream machine-learning models for risk prediction. The random forest algorithm [22] is a supervised learning algorithm constructing an ensemble of decision trees. In this study, we used the Gini index [23] to determine the best predictive variable and location for each tree split in our algorithm. We used a cost complexity parameter to penalize more complex trees and controlled the size of the final tree. The optimal value of the complexity parameter was determined using 5-fold cross-validation. The deep-learning approach is based on the construction of an artificial neural network [24,25], and we trained this method end-to-end by stochastic gradient descent with back propagation. Gradient boosting machines employ a boosting ensemble method by minimizing an exponential loss function of the misclassification rate [26]. Gradient boosting machine performs optimization in the function space by seeking the learner (eg, decision tree) with the maximal negative gradient for the loss function [27,28].

The dataset was iterated 500 times in the model (500 epochs for deep learning, and 500 decision trees for the random forest and gradient boosting machine). A range of values for each hyperparameter was specified and all possible combinations of the hyperparameters were examined; the combination with the highest cross-validation performance metric was obtained. The random forest includes hyperparameters specifying the number of trees and the maximum depth of each tree. The parameters for deep learning included activation, hidden layer size, L1 and L2 regularization, and input dropout ratio as hyperparameters. For gradient boosting machine, a grid search for model optimization was conducted with the maximum number of models, maximum depth of each tree, learning rate, row sample rate per tree, and column sample rate as hyperparameters.

We randomly selected 70% of the total participants to form the training dataset and the remaining 30% were treated as a testing dataset. The training dataset was used for machine learning while the testing dataset was used for assessment of prediction performance of the fully trained classifiers. Five-fold cross-validation was conducted based on the training dataset.

Model Comparisons

The AUC value was adopted to evaluate the performance of the logistic regression and machine-learning models at the predefined time points (3, 5, 7, 10 years). AUC is a robust benchmark model comparison metric for classification models, quantifying the probability of a classifier to differentiate a random positive observation over a random negative observation. The root mean square error was used to verify the result. All analyses were performed using R 3.4.1 statistical software (R Foundation for Statistical Computing, Vienna, Austria), with machine learning toolbox h2o v 3.16.0.2 (H2O.ai Inc, CA, USA). We ranked the top 10 strongest contributing factors to diabetes incidence in all four models.

The relative importance of the risk factors was ranked by their contributions to the variance in the onset of diabetes. For logistic regression, the variance was equal to the squared standardized beta coefficients. For random forest, the variance was the total decrease in node impurities from splitting on the variable, averaged over all trees. For gradient boosting machine, importance was calculated and averaged for each decision tree based on the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for. For deep learning, importance was determined by identifying all weighted connections between the nodes of interest.

Model Prediction

We used the most accurate (highest AUC value) validated model to identify the potential reduction in the probability of diabetes onset by assuming hypothetical changes in participants' BMI categories. We investigated three scenarios: (1) all individuals in the "obese" BMI category (≥30) became "overweight" (BMI=25.0-29.9); (2) in addition to scenario 1, all individuals in the "overweight" BMI category moved to the "healthy" BMI (18.5-24.9) category; and (3) all individuals in the "obese" and "overweight" BMI categories moved to the "healthy" BMI category.

Results

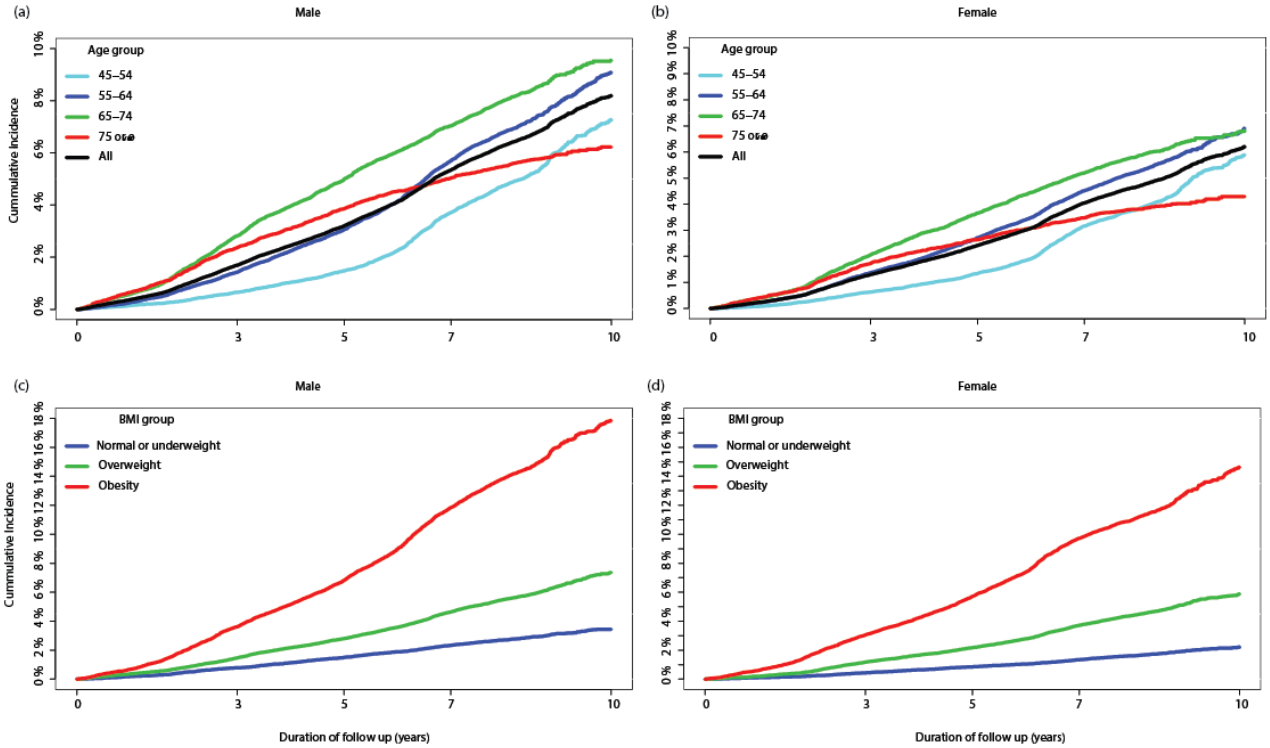
Participant Characteristics

The baseline demographic characteristics of the study population are summarized in Multimedia Appendix 3. In brief, of the 236,684 individuals included in this retrospective cohort study, approximately 6.05% (14,313/236,684) developed T2DM during an average follow up of 8.8 years (range 7.0-11.5; 2,006,194 person years). Individuals with diabetes were significantly more likely to be older, male, overweight or obese, less educated, have a family history of diabetes, reside in a major city, and have a lower income and socioeconomic status (Chi square tests, all P<.0001). Further, individuals with diabetes were more likely to have self-reported hypertension, cardiovascular disease, and dyslipidemia at enrolment (all P<.0001). In terms of lifestyle factors, individuals with diabetes were significantly more likely to be former or current smokers, engage in less physical activity, have longer daily sitting times, consume more processed meat, and have lower milk intake (all P<.0001).

Cumulative Diabetes Incidence by Gender, Age, and BMI Groups

The cumulative incidence of diabetes was significantly higher in men than in women (Figure 1). At the end of 10 years, the cumulative diabetes incidence was 7.66% (7.23%-8.12%) in men, which was significantly higher than that of women (5.84%, range 5.49%-6.20%; odds ratio 1.37, 95% CI 1.32-1.41).

Figure 1. Cumulative incidence of diabetes, stratified by age groups in men and women, and stratified by BMI groups in men and women.



In both men and women, the age group 65-74 years had the highest cumulative incidence of diabetes (10-year incidence: 9.32%, range 8.34%-10.42%), followed by the age groups 45-54 (6.37%, range 5.67%-7.16%), 55-64 (8.68%, range 7.87%-9.57%), and ≥75 (5.84%, range 4.95%-6.88%) years. The incidence of diabetes among participants aged ≥75 years increased at a much slower rate than that of the other age groups and showed a notable reduction after 6-7 years of follow up. This occurred at a time point where the older age group approached the average life expectancy (84.6 years old) in the Australian population [29].

Men with obesity had the highest incidence of diabetes, with a 3, 5, 7, and 10 years cumulative incidence of 3.61% (3.36%-3.89%), 6.82% (6.47%-7.19%), 11.84% (11.37%-12.32%), and 17.39% (15.87%-19.05%), respectively. These were significantly higher than the cumulative incidence in men with a BMI in the overweight and healthy ranges. In particular, the 10-year diabetes incidence in men with obesity was 2.76 (2.61-2.91) and 5.83 (5.41-6.28) higher than that in overweight and healthy weight men, respectively. Diabetes incidence rates in women followed a similar pattern (Figure 1).

Prediction of Diabetes Risk With Machine-Learning Techniques

Machine-learning approaches demonstrated an overall superior prediction of diabetes risk than the conventional regression

analysis (Table 1). The gradient boosting machine model produced the highest accuracy of all four models for 3-year risk prediction. This was followed by the random forest and deep-learning models. Performance measured by AUC in all three machine-learning models was significantly higher than that of the regression analysis (Delong test, all $P<.0001$). A similar pattern was observed for other follow-up durations, but the power of model prediction was reduced by 5%-6% at 10-year follow up. The root mean square error was also the lowest for the gradient boosting machine model (Figure 2, Table 1).

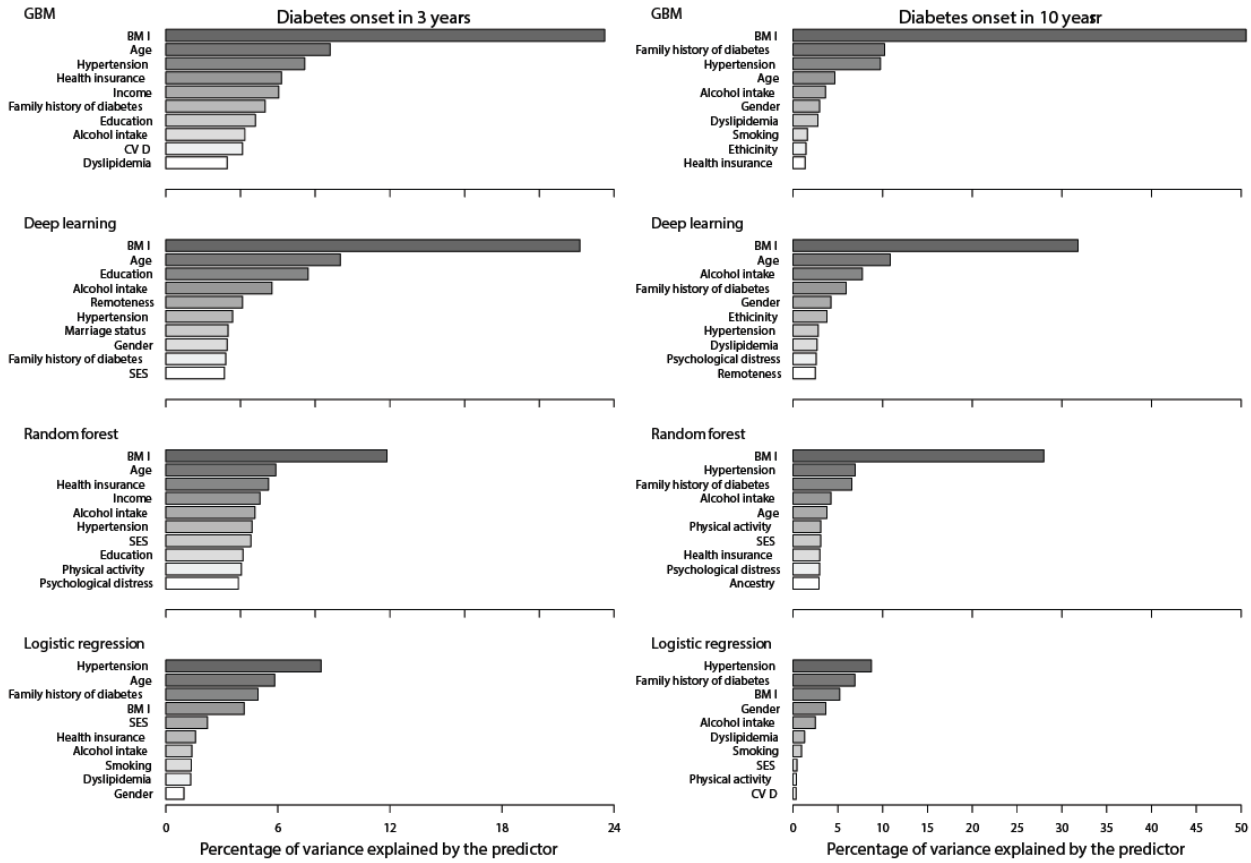
Table 1. Comparison of model performance between logistic regression and machine-learning models.

Duration	Logistic regression		Gradient boosting machine		Deep learning		Random forest	
	AUC[a] (range)	RMSE[b]	AUC (range)	RMSE	AUC (range)	RMSE	AUC (range)	RMSE
3 years	0.7401 (0.7262-0.7541)	0.1203	0.7927 (0.7803-0.8051)	0.1197	0.7769 (0.7639-0.7899)	0.1244	0.7868 (0.7742-0.7993)	0.1198
5 years	0.7192 (0.7084-0.7301)	0.1633	0.7769 (0.7673-0.7864)	0.1620	0.7610 (0.7566-0.7762)	0.1667	0.7769 (0.7612-0.7804)	0.1622
7 years	0.6990 (0.6901-0.7077)	0.2087	0.7589 (0.751-0.7668)	0.2063	0.7526 (0.7446-0.7606)	0.2099	0.7531 (0.7452-0.761)	0.2066
10 years	0.6885 (0.6801-0.6961)	0.2318	0.7491 (0.7426-0.7570)	0.2314	0.7374 (0.7339-0.7486)	0.2435	0.7439 (0.7365-0.7510)	0.2318

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

[b]RMSE: root mean squared error.

Figure 2. Ranked contribution to the variance of diabetes prediction by various models. (+ increasing risk; - decreasing risk; * being male increases risk compared with being female; # being born overseas increases diabetes risk compared with being born in Australia; § having private insurance decreases risk compared with having no private insurance; $ being in major cities increases risk compared with being in inner or outside regional areas; ‡ having Asian or other ancestry increases diabetes risk compared with having Australian ancestry). GBM: gradient boost machine.



The machine-learning models indicated that BMI was the most important predicting factor for the occurrence of diabetes (Figure 2). In the short term (3-year follow up), all three machine-learning models consistently demonstrated that BMI alone contributed to 12%-24% of the variance in the prediction of diabetes. In contrast, BMI contributed to 20%-50% of the variance in the long term (10-year follow up).

Prediction of Diabetes Risk Reduction

Given that BMI was the most important predictor of diabetes, we explored the potential impacts of BMI reduction on the risk

of diabetes onset using the validated gradient boosting machine model. The model predicted that the probability of an obese individual developing diabetes over a 10-year period was approximately one in seven (13.4%, Table 2). In simulated scenario 1, a change of BMI level from obese to overweight significantly reduced the probability of diabetes onset to 6.2% (Table 2). Further, if both obese and overweight individuals were to improve their BMI by a single category (scenario 2), the 10-year probability of diabetes reduced from 8.3% (pooled overweight and obese subgroup) to 3.9%. A greater decline was observed when overweight and obese individuals improved their BMI to the healthy range (scenario 3), with a 10-year probability of diabetes of 2.8%.

Table 2. Model-predicted probability of diabetes onset in three scenarios compared with their respective status quo scenarios.

Scenario	Baseline scenario	Scenarios with hypothetical BMI change	*t* statistic (df)	*P* value
Scenario 1[a](N=46,645)				
Year 3	3.04%	1.54%	6611.97 (93,288)	<.001
Year 5	5.81%	2.89%	7957.43 (93,288)	<.001
Year 7	10.62%	4.68%	12,120.59 (93,288)	<.001
Year 10	13.43%	6.22%	12,732.71 (93,288)	<.001
Scenario 2[b](N=133,830)				
Year 3	1.93%	1.02%	15,401.27 (267,658)	<.001
Year 5	3.68%	1.94%	17,086.55 (267,658)	<.001
Year 7	6.41%	2.98%	23,460.63 (267,658)	<.001
Year 10	8.26%	3.93%	24,604.81 (267,658)	<.001
Scenario 3[c](N=133,830)				
Year 3	1.93%	0.77%	20,856.85 (267,658)	<.001
Year 5	3.68%	1.50%	22,630.22 (267,658)	<.001
Year 7	6.41%	2.14%	31,002.83 (267,658)	<.001
Year 10	8.26%	2.79%	33,214.27 (267658)	<.001

[a]Scenario 1: “obese” individuals but become “overweight.”

[b]Scenario 2: “obese” individuals become “overweight” and “overweight” individuals reach a “healthy” BMI.

[c]Scenario 3: all “obese” and “overweight” individuals reach a “healthy” BMI.

Model Sensitivity and Specificity

We identified the sensitivity and specificity trend versus the risk of diabetes (Multimedia Appendix 4). The trend curves were characterized by a sharp decline in sensitivity and an increase in specificity as the risk of diabetes increased. Crossing of the sensitivity and specificity represents the situation where the two indicators were equal. The model-assigned cut-off levels were consistently lower than the crossing values of the curves, indicating that the models had preferentially weighted on higher sensitivity than specificity.

Discussion

Principal Findings

Our study is a retrospective cohort study of more than 230,000 Australians over a follow-up period spanning a decade. Several important findings can be highlighted. First, we confirmed that machine-learning models performed significantly better than the conventional regression model in predicting the risk of diabetes onset. Notably, the models were developed based solely on self-reported information that was ascertained at a single time point but still achieved 73%-80% accuracy for diabetes prediction for up to 10 years. Second, all machine-learning models consistently demonstrated that BMI is a key risk factor contributing to the onset of T2DM.

Based on these results, we argue that a sophisticated machine-learning model is key for the risk prediction of T2DM onset. In our study, machine-learning models were demonstrated to be superior to the conventional regression model in diabetes risk prediction in a large population-based dataset. Further, the fact that our models were completely based on self-reported information in the absence of any biomarkers suggests the potential for self-assessment in individuals and primary surveillance of diabetes risk in the community. The model tracked over 230,000 Australian individuals for a duration of 10 years and is able to estimate the risk of diabetes development for each individual. Notably, the 10 strongest contributing factors explained over 74%-89% of the variation in diabetes risk. Compared with similar models that are also based on self-reported information [30,31], our model performed consistently better in predicting the risk of diabetes in both the short and long term. This provides further evidence that a simple and user-accessible self-assessment tool can be developed to project the risk of diabetes with robust accuracy, without the assistance of health care workers or need for biomarker sampling or measurement. On a population level, by using a big data

platform, the collection of individual assessment surveys may inform the trends in the diabetes epidemic. This can potentially form an inexpensive user-driven online surveillance platform that surveys diabetes risk factors in a large population, which can in turn forecast the trend of the incidence of diabetes. This is potentially more advantageous than the passive hospital-based case report of diabetes diagnosis that inevitably falls behind the epidemic and population studies that are expensive and unsustainable. Our findings suggest a feasible method such as an electronic health platform for both self-assessment of diabetes risk in individuals and the monitoring of diabetes trends on a population level.

Our finding that BMI is the leading risk factor for T2DM risk was consistent across all machine-learning models. A previous study demonstrated that excessive BMI gain and earlier onset of overweight/obesity are associated with impaired glucose tolerance and diabetes onset [32]. Mokdad et al [33] further demonstrated that being overweight increases the risk of diabetes by 2 fold, while obesity increases the diabetes risk by 3-7 fold. Consistent with previous reports [34], we found that BMI alone accounted for 25%-50% of the variance in diabetes risk.

We further quantified the impact of BMI reduction on the risk of diabetes onset in several hypothetical scenarios. We predicted that reducing an individual's BMI from "obese" to "overweight" would reduce their risk of diabetes in the short and long term by more than half. Further, if BMI could be changed from the "obese and overweight" to "healthy" range, the corresponding risk of diabetes could be reduced by almost two-thirds. This implies that interventions for diabetes prevention should prioritize weight control, especially for those in their late 60s and early 70s. According to the World Health Organization (WHO) global status report on noncommunicable diseases [35], 39% and 12.9% of adults aged 18 years or over in 2014 globally were overweight and obese, respectively, and the worldwide prevalence of obesity has doubled since 1980. Actions to address overweight and obesity are critical to preventing T2DM, as advocated in the WHO report on diabetes [2]. The WHO Global NCD Action Plan 2013–2020 listed halting the rise in diabetes and obesity as one of its voluntary global targets [36]. Our findings are in line with these WHO reports and support their key recommendations.

Strengths and Limitations

The key strengths of the current study include the utilization of a large cohort study dataset (>230,000 participants) with a long follow-up period, and the robust performance of our algorithm for diabetes risk prediction using machine-learning models. Several study limitations should also be noted. First, the analysis was based on a large population survey with information that is subject to self-report bias. Second, the incidence of diabetes in our study was not based on the actual diagnosis of diabetes but was instead inferred by the new use of diabetes-related medications as reported in the Pharmaceutical Benefits Scheme database. This may have resulted in not identifying participants with early diabetes or prediabetes that were not on diabetic medications, and could have therefore underestimated the true diabetes incidence rate over the follow-up period. Nevertheless, one study based on 45 and Up data and linked clinical data proved that diabetes classification based on the Pharmaceutical Benefits Scheme database is more accurate than clinical data [21]. Third, questions related to eating habits in the 45 and Up Study were oversimplified and may not be comparable to standard nutritional surveys. We did not find any association between eating habits and diabetes in our study. Fourth, the absence of mortality data in our dataset implies that the T2DM risk in participants who died before its onset cannot be determined. Fifth, similar to other machine-learning algorithms, the gradient boosting machine model is likely to suffer from overfitting as it automatically removes less fit simulations during its optimization. Regularization parameters and processes such as grid search-tuned learning rate and cross-validation were utilized in this study to enhance the generality of the model. Future work will focus on further validating this model in an independent existing dataset before its official deployment.

Conclusion

In conclusion, we have presented a sophisticated and accurate machine-learning model that allows for the prediction of T2DM incidence for up to 10 years following a single self-reported survey. The model findings highlight the significant impact of higher BMI on diabetes risk and reinforce interventions for weight control to reduce the growing prevalence of diabetes.

Acknowledgments

ME receives support from the University of Melbourne at Research Accelerator Program and the Centre for Eye Research Australia (CERA) Foundation. The CERA receives Operational Infrastructure Support from the Victorian State Government. This specific project is funded by the Australia China Research Accelerator Program at CERA. MH is also supported by the Fundamental Research Funds of the State Key Laboratory in Ophthalmology, National Natural Science Foundation of China (81420108008). The sponsor or funding organization had no role in the design or conduct of this research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. LZ is supported by the National Natural Science Foundation of China (Grant number: 81950410639); Outstanding Young Scholars Funding (Grant number: 3111500001); Xi'an Jiaotong University Basic Research and Profession Grant (Grant number: xtr022019003, xzy032020032); Epidemiology modeling and risk assessment (Grant number: 20200344) and Xi'an Jiaotong University Young Talent Support Grant (Grant number: YX6J004). This research was completed using data collected through the 45 and Up Study and supplied by the Department of Human Services. The 45 and Up Study is managed by the Sax Institute in collaboration with the major partner Cancer Council NSW, and the following partners: the National Heart Foundation of Australia (NSW Division), NSW Ministry of Health, NSW

Government Family & Community Services–Ageing, Carers and the Disability Council NSW, and the Australian Red Cross Blood Service. We thank the many thousands of people participating in the 45 and Up Study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Flowchart for population selection.
[DOCX File , 25 KB - medinform_v8i7e16850_app1.docx]

Multimedia Appendix 2
List of codes for hypoglycemic medications in the Pharmaceutical Benefit Scheme.
[DOCX File , 17 KB - medinform_v8i7e16850_app2.docx]

Multimedia Appendix 3
Demographic, medical and family history, lifestyle and dietary indicators for 236,584 participants in the 45 and Up Study.
[DOCX File , 21 KB - medinform_v8i7e16850_app3.docx]

Multimedia Appendix 4
Sensitivity and specificity trend versus the risk of diabetes by the logistic regression, deep-learning, gradient boosting machine, and random forest models.
[DOCX File , 750 KB - medinform_v8i7e16850_app4.docx]

References

1. Cho N, Shaw J, Karuranga S, Huang Y, da Rocha Fernandes J, Ohlrogge A, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 2018 Apr;138:271-281. [doi: 10.1016/j.diabres.2018.02.023] [Medline: 29496507]
2. Bommer C, Sagalova V, Heesemann E, Manne-Goehler J, Atun R, Bärnighausen T, et al. Global Economic Burden of Diabetes in Adults: Projections From 2015 to 2030. Diabetes Care 2018 May 23;41(5):963-970. [doi: 10.2337/dc17-1962] [Medline: 29475843]
3. Glümer C, Carstensen B, Sandbaek A, Lauritzen T, Jørgensen T, Borch-Johnsen K, Inter99 study. A Danish diabetes risk score for targeted screening: the Inter99 study. Diabetes Care 2004 Mar 26;27(3):727-733. [doi: 10.2337/diacare.27.3.727] [Medline: 14988293]
4. Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. Diabetes Metab Res Rev 2000;16(3):164-171. [doi: 10.1002/1520-7560(200005/06)16:3<164::aid-dmrr103>3.0.co;2-r] [Medline: 10867715]
5. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care 2003 Mar 01;26(3):725-731. [doi: 10.2337/diacare.26.3.725] [Medline: 12610029]
6. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Möhlig M, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care 2007 Mar 27;30(3):510-515. [doi: 10.2337/dc06-2089] [Medline: 17327313]
7. Rahman M, Simmons RK, Harding A, Wareham NJ, Griffin SJ. A simple risk score identifies individuals at high risk of developing Type 2 diabetes: a prospective cohort study. Fam Pract 2008 Jun 30;25(3):191-196. [doi: 10.1093/fampra/cmn024] [Medline: 18515811]
8. Agarwal G, Jiang Y, Rogers Van Katwyk S, Lemieux C, Orpana H, Mao Y, et al. Effectiveness of the CANRISK tool in the identification of dysglycemia in First Nations and Métis in Canada. Health Promot Chronic Dis Prev Can 2018 Feb;38(2):55-63. [doi: 10.24095/hpcdp.38.2.02] [Medline: 29443485]
9. Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. Diabet Med 2010 Aug;27(8):887-895. [doi: 10.1111/j.1464-5491.2010.03037.x] [Medline: 20653746]
10. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, et al. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med 2009 Dec 01;151(11):775-783 [FREE Full text] [doi: 10.7326/0003-4819-151-11-200912010-00005] [Medline: 19949143]
11. Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, et al. AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust 2010 Feb 15;192(4):197-202. [Medline: 20170456]

12. Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. Lancet Diabetes Endocrinol 2014 Jan;2(1):19-29 [FREE Full text] [doi: 10.1016/S2213-8587(13)70103-7] [Medline: 24622666]
13. Hippisley-Cox J, Coupland C. Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. BMJ 2017 Nov 20;359:j5019 [FREE Full text] [doi: 10.1136/bmj.j5019] [Medline: 29158232]
14. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
15. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 2017;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]
16. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data 2015 Dec;3(4):277-287. [doi: 10.1089/big.2015.0020] [Medline: 27441408]
17. Anderson JP, Parikh JR, Shenfeld DK, Ivanov V, Marks C, Church BW, et al. Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records. J Diabetes Sci Technol 2015 Dec 20;10(1):6-18 [FREE Full text] [doi: 10.1177/1932296815620200] [Medline: 26685993]
18. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee Y, et al. Screening for prediabetes using machine learning models. Comput Math Methods Med 2014;2014:618976-661898. [doi: 10.1155/2014/618976] [Medline: 25165484]
19. Worachartcheewan A, Nantasenamat C, Prasertsrithong P, Amranan J, Monnor T, Chaisatit T, et al. Machine learning approaches for discerning intercorrelation of hematological parameters and glucose level for identification of diabetes mellitus. EXCLI J 2013;12:885-893 [FREE Full text] [Medline: 27092034]
20. 45Up Study Collaborators, Banks E, Redman S, Jorm L, Armstrong B, Bauman A, et al. Cohort profile: the 45 and up study. Int J Epidemiol 2008 Oct 19;37(5):941-947 [FREE Full text] [doi: 10.1093/ije/dym184] [Medline: 17881411]
21. Comino EJ, Tran DT, Haas M, Flack J, Jalaludin B, Jorm L, et al. Validating self-report of diabetes use by participants in the 45 and Up Study: a record linkage study. BMC Health Serv Res 2013 Nov 19;13(1):481 [FREE Full text] [doi: 10.1186/1472-6963-13-481] [Medline: 24245780]
22. Breiman L. Random Forests. Mach Learning 2001;45(1):5-32.
23. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.
24. Rosenblatt F. Two theorems on statistical separability in the perceptron. London: Her Majesty's Stationery Office; 1958.
25. Widrow B. Generalization and information storage in networks of Adaline 'Neurons', in Self-Organizing Systems, symposium proceedings. Washington DC: Spartan Books; 1962.
26. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci 1997 Aug;55(1):119-139. [doi: 10.1006/jcss.1997.1504]
27. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001 Oct;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
28. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21. [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
29. Huo L, Shaw JE, Wong E, Harding JL, Peeters A, Magliano DJ. Burden of diabetes in Australia: life expectancy and disability-free life expectancy in adults with diabetes. Diabetologia 2016 Jul 14;59(7):1437-1445. [doi: 10.1007/s00125-016-3948-x] [Medline: 27075450]
30. Habibi S, Ahmadi M, Alizadeh S. Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining. Glob J Health Sci 2015 Mar 18;7(5):304-310. [doi: 10.5539/gjhs.v7n5p304] [Medline: 26156928]
31. Meng X, Huang Y, Rao D, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 2013 Feb;29(2):93-99 [FREE Full text] [doi: 10.1016/j.kjms.2012.08.016] [Medline: 23347811]
32. Power C, Thomas C. Changes in BMI, duration of overweight and obesity, and glucose metabolism: 45 years of follow-up of a birth cohort. Diabetes Care 2011 Sep 20;34(9):1986-1991 [FREE Full text] [doi: 10.2337/dc10-1482] [Medline: 21775760]
33. Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, et al. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. JAMA 2003 Jan 01;289(1):76-79. [doi: 10.1001/jama.289.1.76] [Medline: 12503980]
34. Australian IOH. Australian Burden of Disease Study Impact and causes of illness and death in Australia. Canberra: Australian Institute of Health and Welfare; 2011.
35. World Health Organization. Global status report on noncommunicable diseases. Geneva: World Health Organization; 2014.
36. World Health Organization. Global action plan for the preventioncontrol of NCDs 2013-2020. Geneva: World Health Organization; 2020.

Abbreviations

AUC: area under the curve
NSW: New South Wales
T2DM: type 2 diabetes mellitus
WHO: World Health Organization

Edited by G Eysenbach; submitted 30.10.19; peer-reviewed by Z Ge, L Zhang; comments to author 06.12.19; revised version received 20.02.20; accepted 26.02.20; published 28.07.20.

Please cite as:
Zhang L, Shang X, Sreedharan S, Yan X, Liu J, Keel S, Wu J, Peng W, He M
Predicting the Development of Type 2 Diabetes in a Large Australian Cohort Using Machine-Learning Techniques: Longitudinal Survey Study
JMIR Med Inform 2020;8(7):e16850
URL: https://medinform.jmir.org/2020/7/e16850
doi:10.2196/16850
PMID:32720912

Original Paper

A Predictive Model Based on Machine Learning for the Early Detection of Late-Onset Neonatal Sepsis: Development and Observational Study

Wongeun Song[1*], MS; Se Young Jung[1*], MPH, MD; Hyunyoung Baek[1], RN, MPH; Chang Won Choi[2], MD, PhD; Young Hwa Jung[2], MD; Sooyoung Yoo[1], PhD

[1]Healthcare ICT Research Center, Office of eHealth Research and Businesses, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

[2]Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Sooyoung Yoo, PhD
Healthcare ICT Research Center
Office of eHealth Research and Businesses
Seoul National University Bundang Hospital
172 Dolma-ro, Bundang-gu
Seongnam-si, 13620
Republic of Korea
Phone: 82 32 787 8980
Email: yoosoo0@snubh.org

Abstract

Background: Neonatal sepsis is associated with most cases of mortalities and morbidities in the neonatal intensive care unit (NICU). Many studies have developed prediction models for the early diagnosis of bloodstream infections in newborns, but there are limitations to data collection and management because these models are based on high-resolution waveform data.

Objective: The aim of this study was to examine the feasibility of a prediction model by using noninvasive vital sign data and machine learning technology.

Methods: We used electronic medical record data in intensive care units published in the Medical Information Mart for Intensive Care III clinical database. The late-onset neonatal sepsis (LONS) prediction algorithm using our proposed forward feature selection technique was based on NICU inpatient data and was designed to detect clinical sepsis 48 hours before occurrence. The performance of this prediction model was evaluated using various feature selection algorithms and machine learning models.

Results: The performance of the LONS prediction model was found to be comparable to that of the prediction models that use invasive data such as high-resolution vital sign data, blood gas estimations, blood cell counts, and pH levels. The area under the receiver operating characteristic curve of the 48-hour prediction model was 0.861 and that of the onset detection model was 0.868. The main features that could be vital candidate markers for clinical neonatal sepsis were blood pressure, oxygen saturation, and body temperature. Feature generation using kurtosis and skewness of the features showed the highest performance.

Conclusions: The findings of our study confirmed that the LONS prediction model based on machine learning can be developed using vital sign data that are regularly measured in clinical settings. Future studies should conduct external validation by using different types of data sets and actual clinical verification of the developed model.

(JMIR Med Inform 2020;8(7):e15965) doi:10.2196/15965

KEYWORDS

prediction; late-onset neonatal sepsis; machine learning

Introduction

With the developments in the care system of neonate intensive care units (NICUs), the survival rates of very low birth weight infants have greatly increased. However, neonatal sepsis is still associated with most morbidities and mortalities in the NICUs, and 20% of the deaths in infants weighing <1500 g has been reported to be caused by sepsis. Moreover, infants with sepsis are about three times more likely to die compared to those without sepsis [1]. Neonatal sepsis is categorized into early-onset neonatal sepsis occurring within 72 hours of birth and late-onset neonatal sepsis (LONS) occurring between 72 hours and 120 days after birth [1,2]. Early-onset neonatal sepsis is caused by an in utero infection or by vertical bacterial transmission from the mother during vaginal delivery, while LONS is caused not only by vertical bacterial transmission but also by horizontal bacterial transmission from health care providers and the environment.

Sepsis due to group B *Streptococcus*, which is the most common cause of early-onset neonatal sepsis, can be reduced by 80% before delivery, and intrapartum antibiotic prophylaxis is given when necessary. However, in the case of LONS, unlike early-onset neonatal sepsis, there is no specific antibiotic prophylaxis and there is no robust algorithm that can contribute to its early detection in nonsymptomatic newborns [3,4]. A blood culture test is required for the confirmatory diagnosis of LONS, but it takes an average of 2-3 days to obtain blood culture results. Generally, empirical antibiotic treatments are prescribed to reduce the risk of treatment delay. Even if a negative finding is reported for blood culture, antibiotic therapy is prolonged when the clinical symptoms of LONS are manifested because of the possibility of false-negative blood culture results [5,6]. This treatment process results in bacterial resistance, adverse effects due to prolonged antibiotic therapy, and increased medical costs.

Since several studies have analyzed medical imaging data such as computed tomography and magnetic resonance imaging scans and radiographs by using deep learning and machine learning, recent studies have developed prediction models for the early diagnosis of bloodstream infections and symptomatic systemic inflammatory response syndrome in newborns [7-9].

Griffin et al [10,11] presented a method for identifying the early stage of sepsis by checking the abnormal phase of heart rate characteristics. Stanculescu et al [12] applied the autoregressive hidden Markov model to physiological events such as desaturation and bradycardia in infants and predicted the occurrence of an infection by using the onset prediction model. In addition, a model was presented to make predictions by generating a machine learning model based on vital signs or laboratory features recorded in the electronic medical record (EMR) of an infant [13,14]. However, heart rate characteristics can be affected by respiratory deterioration and surgical procedures in addition to sepsis [15] and heart rate characteristics cannot be obtained in patient monitors without an heart rate characteristic index function. The existing prediction models also involved high computational cost, high-resolution data, or laboratory parameters such as complete blood cell count, immature neutrophil to total neutrophil ratio, and polymorphonuclear leukocyte counts.

Studies on machine learning prediction models using EMR data have inherent problems such as high dimensionality and sparsity, data bias, and few abnormal events [16-18]. Previous studies have tried to resolve the abovementioned problems by using several techniques such as oversampling, undersampling, data handling, and feature selection [19-22]. However, the performance of the model that learned processed data by using data augmentation has not significantly improved compared to that of the previous prediction models, and the EMR-based prediction model is still being challenged [17,20,21]. Therefore, by using data from the Medical Information Mart for Intensive Care III (MIMIC-III) database [23], we aimed to apply the feature selection algorithm to develop a machine learning model that reliably predicts LONS by using low sparsity and few scenarios and to examine the feasibility of the developed prediction model by using noninvasive vital sign data and machine learning technology. In addition, we sought to identify clinically significant vital signs and their corresponding feature analysis methods in LONS.

Methods

Data Source and Target Population

In this study, the MIMIC-III database [23], which consisted of Beth Israel Deaconess Medical Center's public data on admission in the intensive care unit, was used as the data source. The use of data from the MIMIC-III database for research was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and Massachusetts Institute of Technology. NICU inpatients in the 2001-2008 MIMIC-III database were selected as the total population, and their data were extracted. The patients were assigned to sepsis and control groups. The sepsis group consisted of patients with diagnostic codes of septicemia, infections specific to the perinatal period, sepsis, septic shock, systemic inflammatory response syndrome, etc, based on the discharge report. The diagnostic record of the MIMIC-III database utilized the International Classification of Diseases, Ninth Revision, Clinical Modification codes 038 (septicemia), 771 (infants specific to the perinatal period), 995.9 (systemic inflammatory response syndrome), or 785.52 (septic shock), including the abovementioned diagnosis.

Identification of the Sepsis Diagnosis Events

Since the diagnosis table of the MIMIC-III database does not contain information on the timing of diagnosis, this information had to be extracted indirectly from the laboratory test order and intervention information to deduce the timing of diagnosis. Generally, positive blood culture results, clinical deterioration, and high C-reactive protein levels are considered as risk factors, and antibiotic treatment is given by aggregating the information on risk factors [3,4,6]. However, in preterm infants, it is difficult to distinguish the normal conditions of the neonatal period from the clinical signs of sepsis, and since the C-reactive protein value could not be obtained from the MIMIC-III database, the timing of sepsis diagnosis was extracted based on the time of blood culture testing and antibiotic prescription. Generally, a positive blood culture result is selected as the gold standard

based on the criteria used to confirm sepsis. However, since the amount of blood samples that can be collected from preterm or very low birth weight infants is very limited, the number of blood cultures was also small. Moreover, false-negative results may occur because of the low sensitivity of the blood culture, prior use of broad-spectrum antibiotics, and incubation time of the neonatal blood culture [24,25]. Therefore, the timing of sepsis diagnosis was extracted based on the time of the administration of the order of broad-spectrum antibiotics, time of antibiotic administration through intravenous routes, and the time of blood culture order. In the MIMIC-III database, the date on which the item of SPEC_TYPE_DESC in the MICROBIOLOGY EVENTS table was marked as BLOOD CULTURE was assigned as the date of blood culture and the date on which the DRUG was broad-spectrum antibiotics and the ROUTE was filled as IV was used as the antibiotic date in the PRESCRIPTION table.

Feature Processing and Imputation

In the machine learning model, the following features were selected: heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, oxygen saturation, respiratory rate, and body temperature. In the MIMIC-III database, the vital sign and laboratory data that can be used as the candidate features of the predictive models were heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, body temperature, oxygen saturation, Glasgow Coma Scale score, white blood cell count, red blood cell count, platelet count, bilirubin level, albumin level, pH, potassium level, sodium level, creatinine level, blood urea nitrogen, glucose level, partial pressure of carbon dioxide, fraction of inspired oxygen, serum bicarbonate levels, hematocrit, tidal volume, mean airway pressure, peak airway pressure, plateau airway pressure, and Apgar score. Among them, the primary vital signs (body temperature, heart rate, respiratory rate, and blood pressure) and oxygen saturation levels were recorded periodically, whereas the utilization of the other measured values were limited because they were not recorded periodically or they were recorded only for specific patients. Therefore, body temperature, heart rate, respiratory rate, blood pressure, and oxygen saturation that can be commonly applied in predictive models were selected as the features. Moreover, these vital signs are usually accessible from the bedside, do not involve laboratory tests, and can be applied in most hospitals. However, although the current value of the vital signs can be intuitively used as input data, irregular observation cycles of the patient can increase its complexity. Hence, in this study, we tried to increase the accuracy of the actual physiological deterioration of the patient by additionally calculating the statistical and current values of the vital signs and comparing and evaluating the performance of the significant statistical values and observation period for each vital sign. The statistical values, vital signs, and processed observation window size used for the generation are shown in Table 1. In this study, we used statistical values, which are used by many EMR-based prediction models and time series analysis. However, Fourier transform analysis, wavelet transform, and spectrum analysis, which are mainly applied in time series, were excluded from this study because they require high-frequency and relatively periodic data to produce significant results.

For the normal distribution for goodness of fit test, the Shapiro-Wilk test was used for <5000 samples and the Kolmogorov-Smirnov test was used for ≥5000 samples. These normality tests were used when selecting the suitable statistical method depending on the family of distributions. For the correlation, Pearson's correlation was used for normally distributed continuous variables; otherwise, Spearman's correlation was used. Entropy was calculated by estimating the probability density function of the variable with Gaussian kernel density evaluation if a normal distribution was not satisfied. Statistical significance was set to .05. The data quality was assessed by missing value filter and three-sigma rule, and the last observation carried forward method was applied for vital signs assessed as not meeting data quality. The last observation carried forward method is similar to the use of vital signs for diagnosis in general clinical practice and has been mainly used as the imputation method of missing values in clinical prediction models. When there was no measured value, the missing value in the data applied zero imputation to show that it was never measured in the prediction model. Zero imputation was conducted if the calculation could not be performed for reasons such as divided by zero after applying the statistical feature processing.

Table 1. Experimental settings of the vital signs, statistical methods, and processed window time. h: hours.

Category	Experimental options
Value of vital signs	Heart rate, respiratory rate, oxygen saturation, systolic blood pressure, mean blood pressure, diastolic blood pressure, body temperature
Statistical method of feature processing	Mean, median, minimum, maximum, standard deviation, skewness, kurtosis, slope, entropy, delta, absolute delta, correlation coefficient, cross-correlation
Processed observation window size	3 h, 6 h, 12 h, 24 h

Feature Selection Algorithms

To increase the model's performance and to exclude statistical feature values with low feature importance, a method that has been used and verified mainly in the existing machine learning was used. The feature selection method and algorithm were selected because of the large sparsity of data used in this study and because the coefficient was not larger than that of the typical data (Table 1). In addition, the feature selection algorithm presented in this study was applied (Figure 1).

Figure 1. Proposed feature selection algorithm.

```
begin
    Split the patients into training and test set
    while (i < length of the category of vital signs):
        Extract the series x of i th vital sign from raw data
        while (j < length of methods):
            x' = GenerateUnivariateSamples(x, methods[j ])
            Generate the univariate logistic model M_ij(x')
            Evaluate the model quailty(F_ij)
            Append the result of evaluation into F_univariate
        end while
    end while
    Select the vital signs and methods from the highest quality
    Get combination sets using selected vital signs and methods
    while (i < the length of combination sets):
        while (j < the length of patients):
            t_min,j = j th patient's admission time
            t_max,j = j th patient's discharge time
            while (t_min,j + window_size *k < t_max,j) :
                Get j th patient's vital signs in the interval [t_min,j + window_size *k, t_min,j + window_size *(k + 1)], X_ijk
                X'_ijk = GenerateMultivariateInput()
                Add X'_ijk to training and test set
            end while
        end while
        Create the candidate model with ith combination set M_i(x'_i)
        Evaluate M_i(x'_i)
        if Quality(M_i(x'_i)) > Qaulity(M_best(x'_best)):
            M_best(x'_best) = M_i(x'_i)
        end if
        Keep the best model M_best
    end while
    Extract the feature set from the best model M_best
end
```

```
GenerateMultivariateInput
Input : X,
Output : X'
begin
    while (t_min + i * step_size < t_max):
        while (j < the length of methods):
            if j th method is cross coefficient or correlation coefficient :
                x_1 = X in the interval [t_min + i * step_size, t_min + (i+1) * step_size]
                x_2 = X in the interval (t_min + (i+1) * step_size, t_min + (i+2) * step_size]
                x'_ij = method(x_1, x_2)
            else
                x = X in the interval [t_min + i * step_size, t_min + (i+1) * step_size]
                x'_ij = method(x_1)
            end if
    X' = {x'_11, x'_12, ..., x'_ij}
    return X'
end
```

```
GenerateUnivariateSamples
Input : X, method
Output : X'
begin
    Slice X into two subgroup X'_1 and X'_2 ,where X'_1 ∩ X'_2 = Ø and |X'_1| = |X'_2| and |X'_1| + |X'_2| = X
    Compute the results of method(X'_1) and method(X'_2)
    X' = X'_1X'_2
    return X'
end
```

In this study, M is the prediction model, x is the feature derived from each vital sign, and F is the performance of the model and the sum of the receiver operating characteristic (ROC) and average precision. In the case of the data in this study, it was difficult to measure the performance of the minor class when the incidence ratio was too low. Therefore, the classification performance of the major and minor classes for the model selection was evaluated at the same time as the sum of the average precision and area under the ROC (AUROC) curve.

When the features were derived from the vital signs, it was limited to the use of only data obtained from the past observation based on the prediction time to prevent any lookahead due to future observations. In addition, to measure the performance of the proposed feature selection algorithm, we compared the methods usually used from each approach of the feature selection techniques. In the filter approach, chi-squared and mutual information gain were selected. In the embedded approach, lasso linear model L1–based feature selection, extra

tree, random forest, and gradient boosting tree–based feature selection were selected. The other principal component analyses were excluded. Principal component analysis is mainly used in a high dimensional space; thus, an additional analysis of the generated features is needed. This means that the direct interpretation power is relatively low in terms of the correlation between the predicted results and the feature importance. In addition, principal component analysis has several disadvantages such as the feature transformation is possible only when all the existing features are contained and high computational cost. Thus, principal component analysis was excluded owing to the above problems. To minimize the differences between the models' coincidence and temporal characteristics, the observation window and feature processing time stamps were used equally and the model was built without data sampling.

Machine Learning Algorithms

For the classification algorithm of LONS prediction, logistic regression, Gaussian Naïve Bayes, decision tree, gradient boosting, adaptive boosting, bagging classifier, random forest, and multilayer perceptron were selected and assessed. These machine learning classifiers were mainly used in supervised learning methods such as linear model, naive Bayes, decision tree, ensemble method, and neural network model. In the case of the deep learning model, the performance variation was large depending on the number of layers and the change in the learning rate, and the amount of data was not enough to train the deep learning model. Thus, the deep learning models were excluded. To evaluate the performance between the feature selection model and the proposed algorithm, 10% of the target population was used as the feature selection data set, 80% as the train set, and 10% as the test set to perform a stratified 10-fold cross-validation. Then, 100 turns of bootstrapping were applied to obtain the confidence interval for the 95% section of the performance indicator. The model performance indicator enabled a detailed evaluation of the imbalanced data performance by using indicators such as accuracy, AUROC curve, area under the precision-recall curve (APRC), positive predictive value, negative predictive value, and the harmonic mean of precision and recall (F1 score).

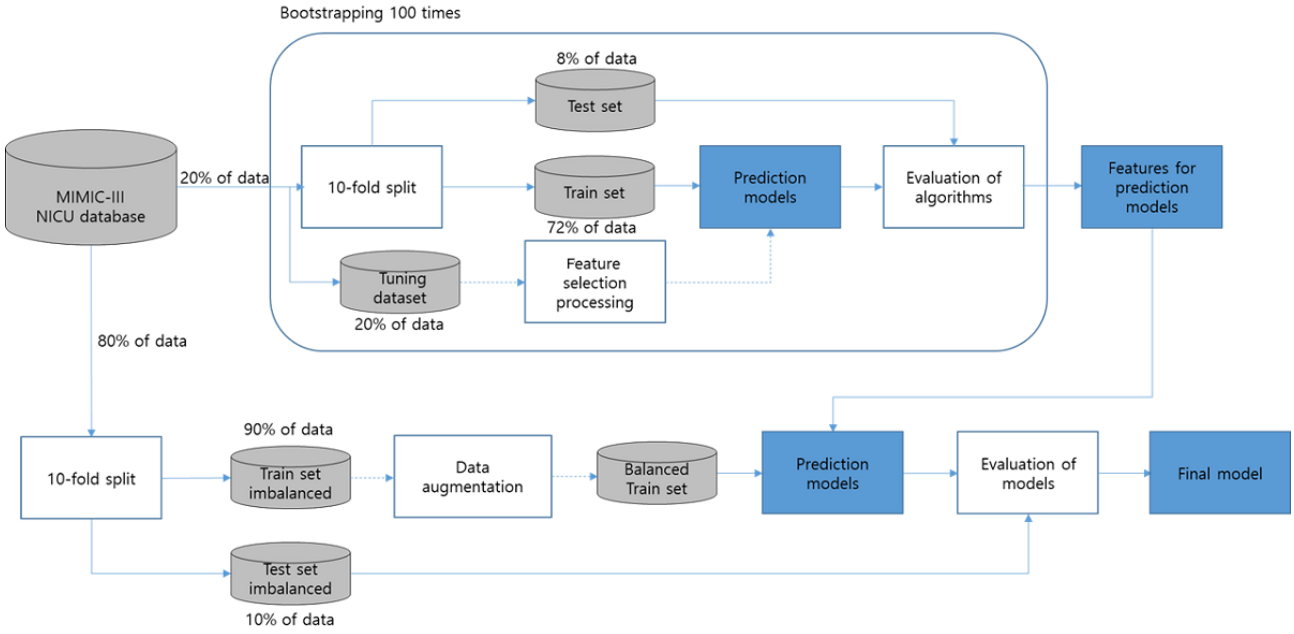
Data Sampling Algorithm

If the data sampling algorithm is applied to model learning after labeling of the data set, a normal model learning is barely attainable because of the imbalanced and overwhelming data. In this study, undersampling algorithms, oversampling algorithms, and a combination of both oversampling and undersampling algorithms, which are data sampling algorithms, were applied to the training set, and the extent to which the model performance for EMR data set was affected was checked using a test set that was not sampled. The oversampling algorithms used were the synthetic minority oversampling technique (SMOTE) [26], adaptive synthetic sampling method [27], and RandomOverSampler. The undersampling algorithms used were NearMiss [28], RandomUnderSampler, All-K-Nearest-Neighbors [29], and InstanceHardnessThreshold [30]. As for the combination of oversampling and undersampling algorithms, SMOTE + Wilson's Edited Nearest Neighbor (SMOTEENN) rule [31] and SMOTE + Tomek links [32] were applied.

Evaluation of the Algorithm

The methods presented in Figure 2 were introduced for the evaluation of the feature selection algorithms and prediction model. To prevent leaking of the test set, the MIMIC-III data were divided by organizing the feature selection evaluation data set at 20% and the prediction model evaluation data set at 80% by using a stratified shuffle. To avoid the overestimation in the test set due to the optimized estimator of 10-fold cross-validation, the performance of the prediction models was measured by initializing the hyperparameters at each fold. For the feature selection algorithm, performance was classified based on the Gaussian Naïve Bayesian Classifier as shown in the study by Phyu and Oo [33]. Given that the classifier's evaluation algorithm is straightforward and that the ensemble classifier such as the gradient-boosted machine can have interactions, nonlinear relationships, and automatically feature selection between features and because there is ambiguity in the statistical properties, the classifier was not selected as the base model [34]. In addition, the mean, minimum, maximum, standard deviation, and median of each vital sign were designated as the baseline features and compared with models that did not perform a feature selection. The existing research model was compared to the model development algorithm presented in this study by presenting both the performance of the presented model and the performance that would have resulted if conducted using the MIMIC-III data. We used Statsmodels and NumPy libraries to analyze the statistical properties. The metric module, a Python module from scikit-learn library, was used to evaluate the classifiers.

Figure 2. Diagram of the evaluation process for models and algorithms. MIMIC-III: Medical Information Mart for Intensive Care; NICU: neonatal intensive care unit.



Results

Characteristics of the Study Population

Table 2 shows the population characteristics of the infants in this study. Of the 7870 infants in the MIMIC-III database, 21 infants were assigned to the clinical LONS group and 2798 infants met the inclusion criteria for the control group. Gestational age, birth weight, and length of stay were significantly different between the clinical LONS and control groups. The median (IQR) gestational age and birth weight in the clinical LONS group were 30 (27.0-34.5) weeks and 0.80 (0.71-1.07) kg, respectively, which were slightly lower than those of the control group whose median (IQR) gestational age and birth weight were 34 (33.5-34.5) weeks and 2.02 (1.58-2.53) kg, respectively. The clinical LONS group showed a significantly longer intensive care unit stay than the control group (87.9 days and 13.3 days, respectively). The male sex rate (%) showed that the male infants in both the clinical LONS and proven sepsis groups had a high risk for infection (61.9% and 51.5%, respectively).

Table 2. Characteristics of the target population (N=7870).

Demographic characteristics	NICU[a], n=96	Clinical LONS[b] group, n=21	Proven sepsis group, n=715	NICU control group, n=2798
Gestational age (week), median (25th-75th percentile)	34.5 (33.5-35.5)	30 (27.0-34.5)	30 (26.6-34.5)	34 (33.5-34.5)
Birth weight (kg), median (25th-75th percentile)	2.56 (0.36-3.27)	0.80 (0.71-1.07)	0.98 (0.72-1.28)	2.02 (1.58-2.53)
Length of stay (day), median (25th-75th percentile)	0.9 (0.1-10.0)	87.9 (61.9-110.9)	71.2 (42.2-107.2)	13.3 (7.1-28.5)
Mortality in the hospital, n (%)	64 (0.8)	3 (3.1)	1 (5.0)	14 (0.5)
Gender, n (%)				
Male, 4243 (53.9)	54 (56.3)	13 (61.9)	368 (51.5)	1508 (53.9)
Female, 3627 (46.1)	42 (43.7)	8 (38.1)	347 (48.5)	1290 (46.1)
Race, n (%)				
White, 4764 (60.5)	56 (58.3)	13 (61.9)	463 (64.8)	1747 (62.4)
African American, 865 (11.0)	14 (14.6)	3 (14.3)	77 (10.8)	301 (10.8)
Asian, 715 (9.1)	2 (2.1)	0 (0.0)	36 (5.0)	161 (5.8)
Hispanic, 369 (4.7)	3 (3.1)	1 (4.8)	29 (4.1)	136 (4.9)
Other, 1157 (14.7)	21 (21.9)	4 (19.0)	110 (15.4)	453 (16.2)
Hospital admission type, n (%)[c]				
Newborn, 7859 (99.9)	95 (99.0)	21 (100.0)	713 (99.7)	2787 (96.4)
Emergency, 220 (2.8)	22 (22.9)	7 (33.3)	9 (1.3)	87 (3.0)
Urgent, 23 (0.3)	0 (0.0)	0 (0.0)	1 (0.1)	16 (0.6)
Elective, 4 (0.1)	0 (0.0)	0 (0.0)	0 (0.0)	2 (0.1)

[a]NICU: neonatal intensive care unit.

[b]LONS: late-onset neonatal sepsis.

[c]allowed to duplicated admission types.

Performance of the Feature Selection Algorithm

The performances of the proposed feature selection algorithm and the existing feature selection algorithm were compared after 100 turns of bootstrapping; the measured performance by the algorithm is shown in Table 3. Given that the AUROC and accuracy rate are likely to be overestimated in the imbalanced data such as this study's data, performance was evaluated based on the APRC and F1 measure, which can evaluate the classification performance for major and minor classes. If the window size is 6 hours, the accuracy of the chi-squared feature selection was the highest at 0.60. The extra tree–based feature selection showed a higher performance with AUROC of 0.79, APRC of 0.23, and F1 score of 0.21. When the goal window size was set at 12 hours, the chi-squared (accuracy 0.68, positive predictive value 0.18), extra tree (APRC 0.24), and the proposed algorithm (AUROC 0.79, F1 score 0.25, and weighted-F1 0.65) showed a higher performance than the baseline. However, the feature selection of the manual information gain and lasso L1 penalty classification was still lower than the performance of the baseline model. In a 24-hour window, the proposed algorithm displayed an overall high performance with AUROC of 0.81 (0.81-0.82), APRC of 0.24 (0.23-0.25), and F1 score of 0.33 (0.32-0.34). When the compatibility interval was evaluated, a uniform performance was displayed despite the variations caused by the sample. Overall, as the duration of the observation window increased, the model receiving the features consisting of statistical values as input had improved performance compared to the baseline feature model. The lasso L1 penalty classification model, which is a univariate method, shows the highest indicator with an accuracy of 0.90. However, an AUROC of 0.69 and F1 score of 0.05 indicate that a feature that can barely distinguish normal from suspected infection conditions was selected. The wrapper method feature selection, which was expected to show a high performance, showed a lower performance than the baseline feature model when the observation window was 6 hours. When the observation time was increased to 12 or 24 hours, the extra tree feature selection showed a high performance. However, as the confidence interval appears wider, the robustness based on the sample population changes is lower than those of the other feature selection algorithms. In particular, the feature selection of the feature importance in the random forest and gradient boosting classifier showed an AUROC of 0.56-0.62 and 0.69-0.75, respectively, at 12 hours, and with the 24-hour window, it showed a wide range of confidence intervals at 0.72-0.79 and 0.75-0.81, respectively.

Table 3. Comparison results for various feature selection algorithms[a].

Window size and algorithm	Accuracy[b], odds ratio (95% CI)	AUROC[c], odds ratio (95% CI)	APRC[d], odds ratio (95% CI)	F1[e], odds ratio (95% CI)	Weighted-F1[f], odds ratio (95% CI)	PPV[g], odds ratio (95% CI)	NPV[h], odds ratio (95% CI)
24 hours							
Proposed	0.76 (0.75-0.78)	*0.81 (0.80-0.81)*	*0.31(0.31-0.32)*	*0.39 (0.38-0.40)*	0.80 (0.79-0.81)	0.28 (0.27-0.29)	0.95 (0.95-0.96)
CS[i]	*0.83 (0.81-0.84)*	0.77 (0.76-0.77)	0.28 (0.27-0.29)	0.34 (0.34-0.35)	*0.83 (0.82-0.85)*	*0.30 (0.29-0.31)*	0.92 (0.92-0.93)
MIG[j]	0.15 (0.13-0.17)	0.53 (0.51-0.54)	0.12 (0.12-0.13)	0.20 (0.20-0.21)	0.08 (0.06-0.11)	0.11 (0.11-0.12)	*0.99 (0.98-0.99)*
LL1[k]	0.27 (0.23-0.31)	0.54 (0.52-0.55)	0.12 (0.12-0.13)	0.22 (0.21-0.22)	0.26 (0.21-0.30)	0.13 (0.12-0.13)	0.93 (0.93-0.94)
ET[l]	0.64 (0.60-0.68)	0.79 (0.77-0.81)	0.31 (0.29-0.32)	0.36 (0.35-0.37)	0.68 (0.64-0.73)	0.24 (0.23-0.26)	0.97 (0.96-0.97)
RF[m]	0.31 (0.27-0.36)	0.65 (0.61-0.68)	0.20 (0.18-0.23)	0.25 (0.24-0.26)	0.30 (0.25-0.36)	0.15 (0.14-0.16)	0.98 (0.98-0.98)
GB[n]	0.49 (0.44-0.54)	0.72 (0.70-0.75)	0.25 (0.23-0.27)	0.30 (0.29-0.32)	0.51 (0.45-0.57)	0.19 (0.18-0.21)	0.97 (0.97-0.98)
Baseline	0.56 (0.53-0.58)	0.77 (0.77-0.77)	0.27 (0.26-0.28)	0.30 (0.29-0.31)	0.62 (0.59-0.65)	0.19 (0.18-0.19)	0.96 (0.96-0.97)
12 hours							
Proposed	0.65 (0.62-0.68)	0.75 (0.75-0.76)	0.25 (0.24-0.25)	*0.31 (0.30-0.32)*	0.70 (0.67-0.73)	0.21 (0.20-0.22)	0.95 (0.95-0.95)
CS	*0.77 (0.75-0.78)*	0.72 (0.71-0.72)	0.22 (0.22-0.23)	0.30 (0.29-0.30)	*0.79 (0.78-0.81)*	*0.23 (0.22-0.23)*	0.93 (0.92-0.93)
MIG	0.17 (0.15-0.19)	0.58 (0.56-0.60)	0.15 (0.14-0.16)	0.21 (0.20-0.21)	0.13 (0.10-0.16)	0.11 (0.11-0.11)	0.98 (0.97-0.98)
LL1	0.11 (0.11-0.11)	0.50 (0.50-0.50)	0.11 (0.11-0.11)	0.20 (0.19-0.20)	0.02 (0.02-0.02)	0.11 (0.11-0.11)	*1.00 (1.00-1.00)*
ET	0.48 (0.44-0.51)	*0.78 (0.77-0.80)*	*0.29 (0.28-0.30)*	0.29 (0.27-0.30)	0.53 (0.49-0.57)	0.17 (0.16-0.19)	0.97 (0.97-0.98)
RF	0.25 (0.21-0.29)	0.61 (0.58-0.64)	0.18 (0.16-0.19)	0.23 (0.22-0.24)	0.22 (0.17-0.27)	0.13 (0.12-0.14)	0.99 (0.99-0.99)
GB	0.41 (0.37-0.45)	0.74 (0.72-0.76)	0.24 (0.23-0.26)	0.27 (0.25-0.27)	0.45 (0.40-0.50)	0.16 (0.15-0.17)	0.97 (0.97-0.98)
Baseline	0.62 (0.59-0.66)	0.73 (0.73-0.74)	0.23 (0.23-0.24)	0.30 (0.30-0.31)	0.67 (0.63-0.71)	0.20 (0.19-0.21)	0.95 (0.95-0.95)
6 hours							
Proposed	0.44 (0.40-0.47)	0.70 (0.70-0.71)	0.20 (0.20-0.21)	0.25 (0.24-0.25)	0.49 (0.45-0.52)	0.15 (0.14-0.16)	0.96 (0.95-0.96)
CS	0.56 (0.52-0.61)	0.67 (0.65-0.68)	0.18 (0.17-0.19)	0.25 (0.24-0.26)	0.60 (0.56-0.65)	0.16 (0.16-0.17)	0.93 (0.93-0.94)
MIG	0.11 (0.11-0.11)	0.50 (0.50-0.50)	0.11 (0.11-0.11)	0.19 (0.19-0.20)	0.03 (0.02-0.03)	0.11 (0.11-0.11)	0.99 (0.98-1.00)
LL1	0.11 (0.11-0.11)	0.50 (0.50-0.50)	0.11 (0.11-0.11)	0.19 (0.19-0.20)	0.02 (0.02-0.02)	0.11 (0.11-0.11)	*1.00 (1.00-1.00)*
ET	0.46 (0.41-0.50)	0.71 (0.69-0.74)	*0.23 (0.22-0.24)*	0.28 (0.26-0.29)	0.49 (0.43-0.54)	0.17 (0.16-0.18)	0.97 (0.96-0.97)
RF	0.30 (0.25-0.34)	0.61 (0.59-0.64)	0.17 (0.15-0.18)	0.17 (0.15-0.18)	0.22 (0.21-0.23)	*0.28 (0.22-0.34)*	0.97 (0.96-0.98)
GB	0.37 (0.32-0.42)	0.66 (0.63-0.69)	0.19 (0.18-0.21)	0.25 (0.24-0.26)	0.38 (0.32-0.44)	0.15 (0.14-0.16)	0.96 (0.95-0.97)
Baseline	*0.60 (0.56-0.63)*	*0.72 (0.71-0.72)*	0.21 (0.21-0.22)	*0.29 (0.21-0.22)*	*0.65 (0.61-0.69)*	0.18 (0.18-0.19)	0.95 (0.95-0.95)

[a]The highest score in each column is shown in italics.

[b]Accuracy: (true positive + true negative) / (positive + negative).

[c]AUROC: area under the receiver operating characteristic.

[d]APRC: area under the precision recall curve.

[e]F1: harmonic mean of precision and recall.

[f]Weighted-F1: macro F1 measurement.

[g]PPV: positive predictive value.

[h]NPV: negative predictive value.

[i]CS: chi-square test.

[j]MIG: mutual information gain.

[k]LL1: lasso L1 penalty classification.

[l]ET: extra tree.

[m]RF: random forest.

[n]GB: gradient boosting.

Performance of Data Sampling

Data sampling was measured by fixing the observation time to 24 hours, applying sampling only on training data using the Gaussian Naïve Bayesian classifier and performing stratified 10-fold cross validation. The results of the accuracy analysis showed that the adaptive synthetic sampling method, All-K-Nearest-Neighbors, InstanceHardnessThreshold, and SMOTEENN performed better than the average value of 0.7, which exceeds the 0.579 of the original data. AUROC and APRC showed that all sampling methods, except SMOTEENN, showed a lower or similar performance to the original ones. In the F1 score, SMOTEENN and instance hardness threshold had a higher performance than the original ones.

Characteristics of the Selected Features

The features obtained from the proposed feature selection method are shown in Table 4. Clinicians might be provided with clinical information on selected features through plots in the form of Table 5 and Multimedia Appendix 1. Table 5 represents the feature importance of the onset after 24 hours calculated by the prediction model learned based on the values of the selected features. Multimedia Appendix 1 provides information on how the prediction model made decisions. Three features were selected among the features mainly selected for each vital sign, and the difference of the latent feature selected based on the window size was confirmed. For the 24-hour window size, the delta between the current and previous measurements was the main variable for all the vital signs. Of these, the kurtosis of the respiratory rate, kurtosis of the body temperature, standard oxygen saturation, and the delta of blood pressures were extracted similarly to the significant feature of the septic shock prediction model [35] for adult patients in the MIMIC-III database, as presented by Carrara et al. As the window size decreased, the data characteristics of the features shifted in importance to mean, entropy, and entropy of delta. This is probably because, in newborns with suspected infection, the frequency of the records increased within the same period such that it affected the entropy increase and was selected as the main variable. When the *P* value of the feature was analyzed using multivariate logistic regression and by focusing on the infection and noninfection points of the statistically significant variables, the oxygen saturation showed desaturation symptoms and wide oxygen saturation changes at the infection point. For the heart rate, tachycardia symptoms were observed at the point of infection. For the body temperature, a delta kurtosis showed a lower expected infection point. Unstable temperature, bradycardia, tachycardia, and hypotension, which are the clinical signs of LONS, were measured [25]. The statistical variable was found to have a lower or similar performance compared to the baseline model for the 12-hour window size. This shows that at least 12 hours of accumulated vital signs must be statistically analyzed so that they can be used as significant physiomarkers.

Table 4. Selected features from the proposed feature selection algorithm.

Vital signs and prediction window size	Statistical method of feature processing
Heart rate	
24 hours	Mean, median absolute delta, minimum absolute delta
12 hours	Mean, minimum absolute delta, median absolute delta
6 hours	Mean, entropy delta, entropy
Respiratory rate	
24 hours	Mean, median absolute delta, kurtosis absolute delta
12 hours	Mean, entropy delta, minimum absolute delta
6 hours	Mean, entropy absolute delta, entropy delta
Oxygen saturation	
24 hours	Mean, standard deviation delta, maximum absolute delta
12 hours	Mean, maximum absolute delta, standard deviation delta
6 hours	Mean, entropy delta, entropy absolute delta
Diastolic blood pressure	
24 hours	Mean, maximum absolute delta, maximum delta
12 hours	Mean, kurtosis delta, kurtosis absolute delta
6 hours	Mean, entropy delta, entropy absolute delta
Mean blood pressure	
24 hours	Mean, maximum absolute delta, maximum delta
12 hours	Mean, maximum absolute delta, kurtosis delta
6 hours	Mean, entropy delta, entropy absolute delta
Systolic blood pressure	
24 hours	Mean, maximum absolute delta, maximum delta
12 hours	Mean, kurtosis delta, kurtosis absolute delta
6 hours	Mean, entropy absolute delta, entropy delta
Body temperature	
24 hours	Mean, kurtosis delta, mean absolute delta
12 hours	Mean, entropy delta, entropy absolute delta
6 hours	Mean, entropy delta, entropy absolute delta

Table 5. An example of the prediction feature importance obtained from the prediction model based on the feature selection algorithm.

Vital signs	Statistical method of feature processing	Feature importance values
Body temperature	Mean	0.282
Oxygen saturation	Mean	0.133
Oxygen saturation	Standard deviation delta	0.126
Heart rate	Mean	0.106
Body temperature	Mean absolute delta	0.052
Heart rate	Median absolute delta	0.046
Respiratory rate	Mean	0.042
Mean blood pressure	Mean	0.032
Body temperature	Kurtosis delta	0.022
Mean blood pressure	Maximum absolute delta	0.022
Diastolic blood pressure	Maximum absolute delta	0.019
Mean blood pressure	Maximum delta	0.018
Respiratory rate	Kurtosis absolute delta	0.017
Systolic blood pressure	Maximum absolute delta	0.016
Diastolic blood pressure	Mean	0.013
Systolic blood pressure	Mean	0.013
Respiratory rate	Median absolute delta	0.011
Oxygen saturation	Maximum absolute delta	0.010
Diastolic blood pressure	Maximum delta	0.009
Systolic blood pressure	Maximum delta	0.006
Heart rate	Minimum absolute delta	0.004

Performance of the Prediction Model

The models presented in this study and those developed in a previous study are shown in Table 6. The following 2 model types were developed based on the onset point: a prediction model that predicts LONS occurrence 48 hours earlier and a detection model that discovers LONS at the time of measurement. The overall performance of the presented model was higher than that of the model presented in previous studies [12,14]. Compared with the NICU sepsis prediction model of MIMIC-III, which has the same data source, the model developed in this study showed a high performance despite the relatively large number of patients. When comparing the model performance, the gradient boosting of the boost type linking multiple week estimators showed an AUROC of 0.881, APRC of 0.536, and F1 score of 0.625 for the prediction model, while the detection model showed a high performance at an AUROC of 0.877, APRC of 0.567, and F1 score of 0.653. The logistic regression and multilayer perceptron with L2 penalty showed an AUROC of 0.874 and 0.860, APRC of 0.558 and 0.496, and F1 scores of 0.593 and 0.542, respectively, for the prediction model, whereas the detection model showed AUROC of 0.874 and 0.860, APRC of 0.558 and 0.534, and F1 scores of 0.615 and 0.595, respectively, which showed an overall higher performance than the existing LONS prediction models.

Table 6. Performance results of the prediction models (microaverage).

Model (Validation data source)	Forecast (h)	Accuracy [a]	AUROC[b]	APRC[c]	F1[d]	Weighted-F1[e]	PPV[f]	NPV[g]
Proposed optimization algorithm LONS[h] prediction model (MIMIC-III)[i]								
Logistic regression	48	0.812	0.861	0.446	0.522	0.835	0.395	0.958
Gaussian Naïve Bayes	48	0.694	0.821	0.394	0.424	0.743	0.283	0.964
Decision tree classifier	48	0.811	0.841	0.449	0.504	0.833	0.389	0.950
Extra tree classifier	48	0.867	0.803	0.367	0.131	0.822	0.527	0.874
Bagging classifier	48	0.863	0.771	0.335	0.251	0.835	0.469	0.883
Random forest classifier	48	0.867	0.805	0.371	0.205	0.831	0.514	0.879
AdaBoost[j] classifier	48	0.825	0.831	0.421	0.507	0.842	0.407	0.944
Gradient boosting classifier	48	0.845	0.859	0.462	0.522	0.856	0.445	0.939
Multilayer perceptron classifier	48	0.811	0.841	0.449	0.504	0.833	0.389	0.950
Proposed optimization algorithm detection model (MIMIC-III)								
Logistic regression	0-48	0.798	0.862	0.568	0.619	0.814	0.501	0.943
Gaussian Naïve Bayes	0-48	0.690	0.806	0.492	0.523	0.720	0.380	0.942
Decision tree classifier	0-48	0.812	0.614	0.306	0.376	0.786	0.572	0.839
Extra tree classifier	0-48	0.809	0.794	0.491	0.180	0.748	0.683	0.813
Bagging classifier	0-48	0.812	0.774	0.461	0.327	0.777	0.592	0.831
Random forest classifier	0-48	0.817	0.825	0.513	0.302	0.775	0.656	0.827
AdaBoost classifier	0-48	0.813	0.835	0.513	0.598	0.822	0.529	0.914
Gradient boosting classifier	0-48	0.830	0.868	0.592	0.624	0.836	0.563	0.919
Multilayer perceptron classifier	0-48	0.799	0.849	0.558	0.611	0.813	0.502	0.935

[a]Accuracy: (true positive + true negative) / (positive + negative).

[b]AUROC: area under the receiver operating characteristic.

[c]APRC: area under the precision recall curve.

[d]F1: harmonic mean of precision and recall.

[e]Weighted-F1: macro-F1 measurement.

[f]PPV: positive predictive value.

[g]NPV: negative predictive value.

[h]LONS: late-onset neonatal sepsis.

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

[j]AdaBoost: adaptive boosting.

Discussion

This study showed that when the biosignals recorded in EMR are used to select and learn features based on the presented algorithm, it is possible to produce a model that can predict LONS 48 hours earlier. Our model also showed a higher or similar performance to the high-resolution model of previous studies. The vital sign–based prediction model, which was based on EMR, showed a model performance that exceeded the model that learned based on the laboratory test, which was presented by Mani et al [14]. When compared with the same classifier, the ROC of the prediction model with our random forest algorithm was 0.805, whereas that of the random forest using the laboratory tests of Mani et al [14] was 0.650, with the vital sign–based learning model showing higher performance. Stanculescu et al's [12] autoregressive hidden Markov model showed an F1 score of 0.690 and APRC of 0.63, which showed higher performance compared to the vital sign–based prediction model that was based on EMR in this study. However, when compared to the detection model, our vital sign–based prediction model that was based on EMR showed a high overall performance. Even if the ROC of the heart rate characteristics was 0.72-0.77, the vital sign–based prediction model recorded in EMR has a higher predictive accuracy than the electrocardiogram-based presentation model [10]. The presented model is expected to show a high contribution even in environments where high-resolution biometric data cannot be collected or where blood culture and laboratory tests cannot be performed regularly. The feature selection presented in this study showed a robust performance compared to the wrapper and embedded method feature selections, which are mainly used in the existing machine learning. Through the selected feature, the main physiomarker can be extracted conversely from EMR. In particular, for preterm infants whose definitions for the

normal range of vital signs are insufficient, statistical variables such as biosignal delta and kurtosis over 24 hours can be used as a basis for classifying a patient's condition. Blood pressure was not used as a key indicator because of the different patient criteria, but it can be used as a major feature by using statistical processing. Moreover, the contribution of the respiratory rate, which was expected to be a key indicator, was low. This is probably because there was a slight change in the respiratory rate of the infants owing to the intervention and ventilation procedures. The correlation coefficient and cross-correlation, which were expected to be important, showed low predictability in low-resolution EMR data. However, they are expected to yield significant results with a high-resolution data set. The vital sign–based prediction model developed in this study has low interpretability, similar to the deep learning and machine learning prediction models in previous studies [12,14]. However, the feature selection presented in this study shows a high performance in linear classifiers such as logistic regression and shows no significant change in performance in other classifiers. If we take advantage of this, applying the feature selection to models such as the fully connected conditional random field and Bayesian inference that have high interpretability can solve the abovementioned problem. Given that the selected feature has dozens of feature spaces, compared to the hundreds of feature spaces in the previous models, simply looking at the model's input variable will have sufficiently high interpretability.

This study has the following limitations. First, external validation is required because the training and test data sets were created within the MIMIC-III database. N-fold cross validation was performed to reduce data bias as much as possible, but the results may vary depending on the clinician's recording cycle, pattern, and policy. Therefore, further research requires progress on whether the model generated by the algorithm is equally applicable to the other EMR databases. Second, the limitation about the data extraction was that the prediction model was generated only with noninvasive signs. This was because the number of noninvasive measurements was relatively higher than that of invasive measurements and thus was extracted from most patients. However, an invasive measurement method has the advantage of providing an accurate measurement value; thus, it is performed for patients requiring intensive observation. In future, it is necessary to study whether there is an improvement in performance when the invasive measurement method is applied to the prediction model of this study. Third, infants without infection may have been included in this study or the timing of sepsis onset might not have been recorded correctly. In clinical practice, empirical antibiotic treatment may be administered to noninfected infants with symptoms of sepsis to reduce mortality. Therefore, there is a limitation that false-positive sepsis can occur. In addition, since the MIMIC-III database covers the period from 2001 to 2008, the data may differ by patient population, treatment, and sepsis definition. Fourth, in the vital sign–based prediction model developed in this study, only multilayer perceptron was applied as a deep learning model. In addition, the performance presented in this study is likely to be lower than the maximum performance that can be modeled because the vital sign–based prediction model that was based on EMR developed in this study is a default model with no hyperparameter tuning. Therefore, advanced deep learning models should be applied to develop sophisticated and accurate prediction models in future studies. Lastly, our model could not be compared with the risk score model and the medical guidelines used in clinical practice. In clinical practice, the results of the hematology tests such as complete blood cell count, immature neutrophil to total neutrophil ratio, and polymorphonuclear leukocyte counts are mainly applied. In the MIMIC-III database used in this study, there was not enough data to record the results of the hematology test as a score model, which makes it difficult to directly compare the performance with the prediction model of the study. Further, ethnicity, gender, and immaturity might affect the outcomes since each factor affects the incidence of sepsis. Previous studies have shown that low birth weight and male gender as risk factors of infection could affect the probability of bloodstream infection. Ethnicity did not seem to directly affect the incidence of sepsis, but the sepsis incidence is different according to the community income level. Therefore, if the aforementioned characteristics of the infants are different from the population of this study, then there is a possibility of obtaining different results. Moreover, the MIMIC-III database lacks the number of infant samples that can be configured for each condition, and it is difficult to show the difference in the results. Nevertheless, acceptable results will be obtained again if the proposed algorithm is reperformed for a specific population. In addition, although the gene type was not recorded in the MIMIC-III database and could not be included, research on gene types should be conducted in the future. If the vital sign–based prediction model that was based on EMR developed in this study is applied to clinical sites, patients with a high LONS risk can be identified up to 48 hours in advance with high accuracy based on the nonregular charts. This could be the basis for triage of patients with a high LONS risk. Combining the predicted results of this algorithm with vital signs traditionally used in clinical sites and test results will help clinicians reach an augmented decision.

In conclusion, we developed a prediction model after generating a key feature with feature selection presented in the EMR data. By doing so, a vital sign–based prediction model that was based on EMR achieved a high prediction performance and robustness compared to the previous feature selection. This research model is expected to significantly reduce the mortality of patients with LONS, and sophisticated predictions can be made through the deep learning model and model optimization. However, the limitations of data extraction and the need to construct a data collection environment remain as the major challenges in applying predictive models in clinical practice. Thus, further research is needed to address these problems.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), which is funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0022).

Conflicts of Interest

None declared.

Multimedia Appendix 1
An example of the decision tree graph for the decision tree classifier.The colors indicate the major class in each node.
[PNG File , 2685 KB - medinform_v8i7e15965_app1.png]

References

1. Hornik C, Fort P, Clark R, Watt K, Benjamin D, Smith P, et al. Early and late onset sepsis in very-low-birth-weight infants from a large group of neonatal intensive care units. Early Human Development 2012 May;88:S69-S74. [doi: 10.1016/s0378-3782(12)70019-1]
2. Stoll BJ, Gordon T, Korones SB, Shankaran S, Tyson JE, Bauer CR, et al. Late-onset sepsis in very low birth weight neonates: a report from the National Institute of Child Health and Human Development Neonatal Research Network. J Pediatr 1996 Jul;129(1):63-71. [doi: 10.1016/s0022-3476(96)70191-9] [Medline: 8757564]
3. Fanaroff AA, Korones SB, Wright LL, Verter J, Poland RL, Bauer CR, et al. Incidence, presenting features, risk factors and significance of late onset septicemia in very low birth weight infants. The National Institute of Child Health and Human Development Neonatal Research Network. Pediatr Infect Dis J 1998 Jul;17(7):593-598. [doi: 10.1097/00006454-199807000-00004] [Medline: 9686724]
4. Bekhof J, Reitsma JB, Kok JH, Van Straaten IHLM. Clinical signs to identify late-onset sepsis in preterm infants. Eur J Pediatr 2013 Apr;172(4):501-508. [doi: 10.1007/s00431-012-1910-6] [Medline: 23271492]
5. Borghesi A, Stronati M. Strategies for the prevention of hospital-acquired infections in the neonatal intensive care unit. J Hosp Infect 2008 Apr;68(4):293-300. [doi: 10.1016/j.jhin.2008.01.011] [Medline: 18329134]
6. Sivanandan S, Soraisham AS, Swarnam K. Choice and duration of antimicrobial therapy for neonatal sepsis and meningitis. Int J Pediatr 2011;2011:712150 [FREE Full text] [doi: 10.1155/2011/712150] [Medline: 22164179]
7. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018 Nov;15(11):e1002683 [FREE Full text] [doi: 10.1371/journal.pmed.1002683] [Medline: 30399157]
8. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2]
9. Malak JS, Zeraati H, Nayeri FS, Safdari R, Shahraki AD. Neonatal intensive care decision support systems using artificial intelligence techniques: a systematic review. Artif Intell Rev 2018 May 22;52(4):2685-2704. [doi: 10.1007/s10462-018-9635-1]
10. Griffin MP, O'Shea TM, Bissonette EA, Harrell FE, Lake DE, Moorman JR. Abnormal Heart Rate Characteristics Preceding Neonatal Sepsis and Sepsis-Like Illness. Pediatr Res 2003 Jun;53(6):920-926. [doi: 10.1203/01.pdr.0000064904.05313.d2]
11. Griffin MP, Moorman JR. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. Pediatrics 2001 Jan;107(1):97-104. [doi: 10.1542/peds.107.1.97] [Medline: 11134441]
12. Stanculescu I, Williams CKI, Freer Y. Autoregressive hidden Markov models for the early detection of neonatal sepsis. IEEE J Biomed Health Inform 2014 Sep;18(5):1560-1570. [doi: 10.1109/JBHI.2013.2294692] [Medline: 25192568]
13. Stanculescu I, Williams C, Freer Y. A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. 2014 Jun 24 Presented at: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence; 23 July 2014; Quebec City, Quebec, Canada.
14. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Inform Assoc 2014;21(2):326-336 [FREE Full text] [doi: 10.1136/amiajnl-2013-001854] [Medline: 24043317]
15. Sullivan BA, Grice SM, Lake DE, Moorman JR, Fairchild KD. Infection and other clinical correlates of abnormal heart rate characteristics in preterm infants. J Pediatr 2014 Apr;164(4):775-780 [FREE Full text] [doi: 10.1016/j.jpeds.2013.11.038] [Medline: 24412138]
16. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 02;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
17. Wu J, Roy J, Stewart WF. Prediction Modeling Using EHR Data. Medical Care 2010;48:S106-S113. [doi: 10.1097/mlr.0b013e3181de9e17]

18. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
19. Rahman MM, Davis DN. Addressing the Class Imbalance Problem in Medical Datasets. IJMLC 2013:224-228. [doi: 10.7763/ijmlc.2013.v3.307]
20. Segura-Bedmar I, Colón-Ruíz C, Tejedor-Alonso M, Moro-Moro M. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. J Biomed Inform 2018 Nov;87:50-59 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.012] [Medline: 30266231]
21. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw 2008;21(2-3):427-436 [FREE Full text] [doi: 10.1016/j.neunet.2007.12.031] [Medline: 18272329]
22. Zheng K, Gao J, Ngiam K, Ooi B, Yip W. Resolving the Bias in Electronic Medical Records. New York, NY, USA: Association for Computing Machinery; 2017 Presented at: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2017; Halifax, NS, Canada p. 2171-2180 URL: https://doi.org/10.1145/3097983.3098149 [doi: 10.1145/3097983.3098149]
23. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
24. Guerti K, Devos H, Ieven MM, Mahieu LM. Time to positivity of neonatal blood cultures: fast and furious? J Med Microbiol 2011 Apr;60(Pt 4):446-453. [doi: 10.1099/jmm.0.020651-0] [Medline: 21163823]
25. Zea-Vera A, Ochoa TJ. Challenges in the diagnosis and management of neonatal sepsis. J Trop Pediatr 2015 Feb;61(1):1-13 [FREE Full text] [doi: 10.1093/tropej/fmu079] [Medline: 25604489]
26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
27. He H, Bai Y, Garcia E, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. : IEEE; 2008 Jun Presented at: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 1-8 June 2008; Hong Kong, China p. A. [doi: 10.1109/IJCNN.2008.4633969]
28. Zhang J, Mani I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. 2003 Aug 21 Presented at: Proceeding of International Conference on Machine Learning (ICML 2003), Workshop on Learning from Imbalanced Data Sets; 21 August 2003; Washington DC.
29. Tomek I. An Experiment with the Edited Nearest-Neighbor Rule. IEEE Trans. Syst., Man, Cybern 1976 Jun;SMC-6(6):448-452. [doi: 10.1109/tsmc.1976.4309523]
30. Smith M, Martinez T, Giraud-Carrier C. An instance level analysis of data complexity. Mach Learn 2013 Nov 5;95(2):225-256 [FREE Full text] [doi: 10.1007/s10994-013-5422-z]
31. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl 2004 Jun 01;6(1):20-29. [doi: 10.1145/1007730.1007735]
32. Batista GEAPA, Bazzan ALC, Monard MC. Balancing Training Data for Automated Annotation of Keywords: a Case Study. In: Brazilian Workshop on Bioinformatics. 2003 Presented at: Proceedings of the Second Brazilian Workshop on Bioinformatics; December 3-5, 2003; Macaé, Rio de Janeiro, Brazil p. 10-18 URL: https://pdfs.semanticscholar.org/c1a9/5197e15fa99f55cd0cb2ee14d2f02699a919.pdf
33. Phyu TZ, Oo NN. Performance Comparison of Feature Selection Methods. MATEC Web of Conferences 2016 Feb 17;42:06002. [doi: 10.1051/matecconf/20164206002]
34. Friedman JH. Stochastic gradient boosting. Computational Statistics & Data Analysis 2002 Feb;38(4):367-378. [doi: 10.1016/s0167-9473(01)00065-2]
35. Carrara M, Baselli G, Ferrario M. Mortality Prediction Model of Septic Shock Patients Based on Routinely Recorded Data. Comput Math Methods Med 2015;2015:761435 [FREE Full text] [doi: 10.1155/2015/761435] [Medline: 26557154]

Abbreviations

APRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic
EMR: electronic medical record
LONS: late-onset neonatal sepsis
MIMIC-III: Medical Information Mart for Intensive Care III
NICU: neonatal intensive care unit
ROC: receiver operating characteristic
SMOTE: synthetic minority oversampling technique
SMOTEENN: SMOTE + Wilson's Edited Nearest Neighbor Rule

Edited by C Lovis; submitted 22.08.19; peer-reviewed by SY Shin, F Agakov, A Aminbeidokhti; comments to author 16.10.19; revised version received 27.03.20; accepted 07.06.20; published 31.07.20.

Please cite as:
Song W, Jung SY, Baek H, Choi CW, Jung YH, Yoo S
A Predictive Model Based on Machine Learning for the Early Detection of Late-Onset Neonatal Sepsis: Development and Observational Study
JMIR Med Inform 2020;8(7):e15965
URL: http://medinform.jmir.org/2020/7/e15965/
doi:10.2196/15965
PMID:32735230

Original Paper

Positioning and Utilization of Information and Communication Technology in Community Pharmacies of Selangor, Malaysia: Cross-Sectional Study

Bhuvan KC[1], BPharm, MPharm, PhD; Dorothy Lim[1], BPharm; Chia Chia Low[1], BPharm; Connie Chew[1], BPharm; Ali Qais Blebil[1], MPharm, PhD; Juman Abdulelah Dujaili[1], MPharm, PhD; Alian A Alrasheedy[2,3], BPharm, MPharm, PhD

[1]School of Pharmacy, Monash University Malaysia, Bandar Sunway, Malaysia

[2]Unaizah College of Pharmacy, Qassim University, Unaizah, Saudi Arabia

[3]Department of Pharmacy Practice, College of Pharmacy, Qassim University, Buraydah, Saudi Arabia

Corresponding Author:
Bhuvan KC, BPharm, MPharm, PhD
School of Pharmacy
Monash University Malaysia
Jalan Lagoon Selatan
Bandar Sunway, 47500
Malaysia
Phone: 60 142271852
Email: bhuvan.kc@monash.edu

Abstract

Background: Information and communication technology (ICT) is an essential element of modern “smart” cities. These smart cities have integrated housing, marketplace, public amenities, services, business, and transportation via ICT. ICT is also now widely used in urban health care delivery.

Objective: The aim of this study was to determine the positioning and roles of ICT in community pharmacies in the state of Selangor, Malaysia.

Methods: A cross-sectional study was conducted from November 2018 to January 2019 across 9 different subdistricts in the state of Selangor, including Subang Jaya, Cheras, Puchong, Port Klang, Kota Kemuning, Selayang, Chow Kit, Ampang, and Seri Kembangan. A total of 90 community pharmacists were approached from the 9 subdistricts and invited to participate in the study.

Results: Of the 90 community pharmacies approached, 60 agreed to participate in the study, representing a response rate of 67%. The majority (36/60, 60%) of the respondents were women, and more than half (32/60, 53%) of the community pharmacies were run by young adults (ie, 30 years old and younger). More than three-quarters of the community pharmacies (46/60, 77%) used electronic health records. Half of the community pharmacies used online social media platforms for advertising and promoting their pharmacies. The vast majority of the community pharmacies (55/60, 92%) were using modern electronic payment systems, and some were also using other new electronic payment methods. Moreover, most of the community pharmacies (41/60, 68%) were using software and programs for accounting and logistics purposes. In addition, 47/60 (78%) of the community pharmacies used a barcode reading system for medicines/health products, and 16/60 (27%) of the pharmacies had online stores, and consumers could buy medicines and health products from these pharmacies via their online portal. In addition, 20/60 (33%) of the community pharmacies used at least one of the common online business platforms available in Southeast Asia to sell products/medicines. The telephone was the most commonly used means of communication with patients, although some pharmacies also used email, WhatsApp, SMS text messaging, and other communication platforms.

Conclusions: This study showed that the majority of community pharmacies in Selangor, Malaysia are using ICT for different purposes. However, there is still limited use of mobile apps to provide health services. Overall, community pharmacies have been adopting ICT apps for pharmacy services but the rate of adoption is relatively slower than that in other sectors of Malaysia.

(JMIR Med Inform 2020;8(7):e17982) doi:10.2196/17982

KEYWORDS

information and communication technology; community pharmacy; Malaysia; pharmacy services

Introduction

Information and communication technology (ICT) is an important element in contemporary business and lifestyle. ICT includes various electronic apps, tools, and services that promote the sharing of information and facilitate communication [1]. Malaysia is undergoing rapid urbanization, with ICT playing an integral role, especially in "smart" cities [2]. Under the concept of smart cities, in which cities are at the interface of the social, economic, and technological dimensions, ICT is used to enhance the quality of life of the population, improve accessibility to services, and ensure consistent improvement in the economy and in sustainable social and environmental developments [3]. ICT apps are used in various activities such as accounting, marketing, staffing, record keeping, manufacturing processes (consumer and industrial goods), and in the service sectors (eg, transportation, education, health care) [2].

The advancement of ICT apps accompanied by reduction in costs and increased availability have resulted in their increased use and improvement along with related services for various functions by both the public and private sectors. The health care sector has also increasingly adopted ICT apps for its various services. Community pharmacies in the countries of the Organization for Economic Cooperation and Development, such as the United States and European countries, are already using ICT for various purposes, including for dispensing medicines, billing, and government reimbursements [4]. Furthermore, ICT apps are being used in community pharmacies to manage various services, including drug information systems, laboratory systems, logistics, and accounting [5]. A futuristic view shows that the adoption of ICT apps in the community pharmacy setting is expected to contribute to clinical decision making, achieving cost-effectiveness, expedited medication delivery times, and faster delivery of many other services [6].

The Malaysian health system is a dual-tiered system with a government-funded public sector and a private sector [7]. Pharmacists are an integral component of the Malaysian health system and play a vital role in regulatory control and policy work for community pharmacies, hospital and clinical pharmacy services, pharmaceutical production, academic activities, and health promotion [8]. Community pharmacists are the first-line health care providers for patients with minor ailments or those seeking health care services and over-the-counter medicines [9]. In Malaysia, community pharmacies operate as independent retail pharmacies, retail chain pharmacies, or pharmacies attached to a medical doctor's clinic; they provide medicines, basic health care advice, and other pharmacy services such as health education and drug information.

Malaysia now has smart cities with integrated housing, marketplaces, public amenities, and transportation linkages that are facilitated and connected via ICT networks and apps [10]. To provide medicines and health services to the people living in these smart cities, community pharmacies need to reposition themselves and adopt ICT apps. Furthermore, the Malaysian community pharmacy sector faces challenges such as pressure to improve productivity and services to meet the needs of patients and consumers and the ever-increasing demand for various kinds of health services. The implementation of ICT in community pharmacies can streamline and help to overcome these challenges, especially for the effective delivery of pharmacy services [11]. However, there is a dearth of information regarding the utilization of ICT by community pharmacies in Malaysia. Therefore, the aim of the present study was to understand the positioning and utilization of ICT by community pharmacies in Malaysia, and to study its impact on the practices of community pharmacies toward realizing better productivity and health outcomes for patients.

Methods

Study Design and Setting

A cross-sectional study was conducted among community pharmacies using a self-administered questionnaire in the state of Selangor, Malaysia from November 2018 to January 2019.

Participants

We obtained a list of the total number of pharmacies in the state of Selangor from the website of the Pharmaceutical Services Division, Ministry of Health, Malaysia. There were 1394 registered community pharmacies identified in the state of Selangor. We used Google Maps to locate pharmacies in 9 subdistricts and selected community pharmacies based on ease of access to obtain a minimum of 5 community pharmacies from each subdistrict. Accordingly, a total of 90 community pharmacies spread across 9 subdistricts of the state of Selangor were approached to participate in this study. In our study, the sampling unit was the community pharmacy. Consequently, we invited only one participant from pharmacies with more than one pharmacist.

Study Instrument/Tool

A questionnaire was developed following a literature review on ICT use in the community pharmacy and health sector. We also obtained local literature on ICT and health in Malaysia to inform the development of the questionnaire [1-8,11]. The questionnaire comprised two parts. Part A encompassed the demographic characteristics of participants and their community pharmacies, and Part B encompassed the roles of ICT in community pharmacies. The questionnaire was finalized via a pilot study with 5 community pharmacists.

Data Collection

The community pharmacies were approached by trained research assistants who invited the pharmacists to participate in the study. The questionnaire, along with a self-explanatory statement, was administered to the community pharmacies, and written consent to participate in the study was obtained. The participants were then asked to complete the questionnaire and were informed that they could complete it at their convenience and that the

questionnaires would be collected after 1 week. The questionnaire was in English, and it was estimated to take 15 to 20 minutes to complete. After 1 week, the research assistants visited the pharmacies and collected the questionnaires. Owing to logistical barriers, no further follow-up visits were made.

Data Analysis

The responses from the paper-based questionnaires were transferred into and analyzed using the Statistical Package for Social Sciences version 20.0 (SPSS Inc, Chicago, IL, USA). Categorical variables are presented as numbers and frequencies, whereas continuous variables are presented as means (SD). The association between sociodemographic characteristics and the adoption of ICT in community pharmacies was examined by the Chi square test and its alternative, Fisher exact test, when relevant. A *P* value of less than .05 was considered to indicate statistical significance.

Ethics Approval

Ethics approval was obtained from the Monash University Human Research Ethics Committee (Project ID: 16602, date of approval October 1, 2018) prior to study commencement.

Results

Among the 90 community pharmacies approached, 60 agreed to participate in the study, for a response rate of 67%. Among the respondents, the majority were female pharmacists, and more than half of the community pharmacies were run by young adults (ie, 30 years old and younger). Most of the respondents held a bachelor's degree in pharmacy (Table 1). Regarding location and accessibility, most of the community pharmacies were located in residential areas, including towns and near markets. Some pharmacies were also located inside shopping malls and near hospitals. Most of the community pharmacies were independent retail pharmacies, followed by chain pharmacies. The majority of the community pharmacies remained open 8-12 hours per day. Community pharmacies provided a range of different services such as blood pressure measurement, blood glucose tests, blood cholesterol tests, and other services (eg, diet consultation, pregnancy tests, weight management, smoking cessation). The detailed results are summarized in Table 1.

As shown in Table 2, almost all of the community pharmacies were locatable via GPS and associated navigation apps such as Waze and Google Maps. Half of the community pharmacies used social media for the advertisement and promotion of their products. The majority of the community pharmacies were using electronic payment systems, including credit cards, and some were also using other new electronic payment methods such as Alipay, Boost, and Epay. Moreover, many of the community pharmacies were using software and programs for accounting and logistics purposes. In addition, the majority of participating community pharmacies were using a barcode reading system for medicines/health products. Overall, 16/60 (22%) of the pharmacies had online stores, and consumers could buy medicines and health products from these pharmacies via their online portal. In addition, a third of those pharmacies were using at least one of the common online business platforms in Southeast Asia to sell products/medicines (Table 2).

Table 1. Demographic and characteristic data of participants and their pharmacies (N=60).

Characteristics	n (%)[a]
Gender	
Male	24 (40)
Female	36 (60)
Age (years)	
21-30	32 (53)
31-40	14 (23)
41-50	9 (15)
51-60	2 (3)
61-70	2 (3)
71-80	1 (2)
Qualification	
Bachelor's degree	38 (63)
Master's degree	8 (13)
Diploma and other technical degree	14 (23)
Location of community pharmacy	
In a shopping mall	7 (12)
Near a residential area (eg, in a town, market)	44 (73)
Near a hospital/clinic	7 (12)
Rural area	2 (3)
Type of community pharmacy	
Independent retail pharmacy	50 (83)
Wholesale outlet	1 (2)
Chain pharmacy	9 (15)
Number of hours open	
8-12	53 (88)
>12	7 (12)
Number of staff members	
<5	32 (53)
5-10	23 (38)
>10	3 (5)
Not reported	2 (3)
Number of prescriptions per day	
5 or less	60 (100)
>5	0 (0)
Number of patients per day	
≤50	26 (43)
50-100	15 (25)
>100	5 (8)
Not reported	14 (23)
Number of medicines dispensed/sold per day	
≤50	20 (33)
51-100	22 (37)

Characteristics	n (%)[a]
>100	6 (10)
Not reported	12 (20)
Health services provided	
Blood pressure measurement	37 (62)
Blood glucose test	34 (57)
Blood lipid test	11 (18)
Blood cholesterol test	25 (42)
Blood uric acid	10 (17)
Other services[b]	25 (42)

[a]Due to rounding, percentages may not add up to 100%.

[b]Other services include diet consultation, pregnancy tests, weight management, and smoking cessation.

Regarding automation and the use of technology in preparing and dispensing medicines, the majority of the community pharmacies (>90%) did not have a tablet-based or pill-counting machine or an automated unit-dose packing machine (Table 2). However, some community pharmacies had a labeling machine. As shown in Table 2, most pharmacies had computers with internet access. In addition, the majority of the pharmacies had an electronic record system and online accounts of the patients. The telephone was the most commonly used means of communication with patients, although some pharmacies also used email, WhatsApp, and text messaging.

With respect to the use of online resources to provide evidence-based medicine information, the majority of the community pharmacies had access to drug information resources online and through search engines. The pharmacies were using globally accessed medical and health-related portals such as Monthly Index of Medical Specialties, Medscape, National Pharmaceutical Regulatory Agency, and the British National Formulary to look at health- and medicine-related information (Table 2).

Statistical analysis was performed to examine whether there were any associations between the sociodemographic characteristics and the utilization of ICT in community pharmacies (Table 3). Several associations were noted. There was a statistically significant association between gender and clients having an online account in the pharmacy (P=.01). A total of 69% (25/36) of females reported having an online account compared with only 38% (9/24) of male respondents.

There were two statistically significant associations found among the associations evaluated: age of respondents was associated with having a labeling machine in the pharmacy and having mobile or online apps for the pharmacy store. Among pharmacists aged ≤30 years, 41% (13/32) reported having a labeling machine compared to only 14% (2/14) and 7% (1/14) of those aged 31-40 and >40, respectively. Similarly, 34% (11/32) of those aged ≤30 years reported having mobile or online apps for pharmacy stores compared to only 7% (2/28) of respondents aged above 30.

There was a statistically significant association between the type of pharmacy and the ability to receive patient information from different mobile health apps (P=.002), with 56% (5/9) of the chain pharmacies reporting receiving information compared to only 12% (6/50) of retail pharmacies. In addition, there was a significant association between the number of staff and having an online store for the pharmacy (P=.009) and receiving patient information from different mobile health apps (P=.01). In this study, 48% (11/23) of pharmacies with 5-10 staff members had an online store, compared to only 33% (1/3) and 13% (4/32) of pharmacies with more than 10 staff members and less than 5 staff members, respectively. In addition, 39% (9/23) of pharmacies with 5-10 staff members reported receiving patient information from different mobile health apps compared to only 6% (2/32) in pharmacies with less than 5 staff members.

Table 2. Utilization of information community technology in community pharmacies (N=60).

Variable	n (%)
Locatable via GPS and other navigation systems	59 (98)
Online advertisement medium	
Facebook	28 (47)
Twitter	1 (2)
Instagram	1 (2)
None	30 (50)
Electronic payment systems	55 (92)
Tablet- or pill-counting machine	5 (8)
Barcode reading system for medicines/health products	47 (78)
Automated/unit-dose packaging machine	3 (5)
Labeling machine	16 (27)
Computer with internet facilities	56 (93)
Electronic patient record system	46 (77)
Online account of the clients	34 (57)
Communication with regular clients	
Phone	50 (83)
Email	12 (20)
WhatsApp	9 (15)
Text message	22 (37)
Walk-in	26 (43)
Facebook	1 (2)
Fax	1 (2)
No communication	1 (2)
Online store of the pharmacy	16 (27)
Mobile or online apps for pharmacy store	13 (22)
Receipt of patient information from different mobile health apps	11 (18)
Website/software for drug information	52 (87)
Common website/software and sources for drug information	
MIMS[a]	41 (68)
Google search	13 (22)
Medscape	2 (3)
National Pharmaceutical Regulatory Agency	3 (5)
Micromedex, Lexicomp	1 (2)
British National Formulary	1 (2)
NHS[b] Health	1 (2)
Up-to-Date	1 (2)
PubMed, NICE[c]	1 (2)
Use of common online business platforms to sell products/medicines	
Lazada	10 (17)
Shopee	6 (10)
Esyms	8 (13)

Variable	n (%)
11street	2 (3)
None	40 (67)
Software/programs for logistics and accounting	41 (68)

[a]MIMS: Monthly Index of Medical Specialties.

[b]NHS: National Health Service.

[c]NICE: National Institute for Health and Care Excellence.

Table 3. Associations between sociodemographic characteristics and adoption of information communication technology in community pharmacies.[a]

Variable	Gender	Age	Qualification	Location of pharmacy	Type of pharmacy	Number of staff
Locatable via GPS and other navigation systems	.40	.46	.90	.58	.65	.83
Online advertisement medium	.37	.02	.57	.17	.37	.86
Electronic payment systems	.99	.75	.44	.16	.62	.28
Tablet or pill-counting machine	.38	.15	.58	.95	.19	.33
Barcode reading system for medicines/health products	.31	.48	.19	.73	.58	.37
Automated/unit-dose packaging machine	.56	.40	.61	.53	.43	.67
Labeling machine	.27	.04	.80	.68	.39	.38
Computer with internet facilities	.67	.80	.64	.10	.13	.92
Electronic patient record system	.38	.29	.73	.29	.43	.18
Online account of the clients	.01	.13	.26	.53	.82	.16
Communication with regular clients	.44	.06	.10	.39	.12	.05
Online store of the pharmacy	.12	.32	.53	.90	.30	.009
Mobile or online apps for pharmacy store	.07	.04	.83	.74	.12	.05
Receipt of patient information from different mobile health apps	.68	.88	.74	.95	.005	.01
Website/software for drug information	.91	.71	.36	.34	.23	.80
Use of common online business platforms to sell products/medicines	.26	.69	.23	.61	.13	.52
Software/programs for logistics and accounting	.64	.58	.21	.40	.39	.31

[a]Values in the table are *P* values from the Chi square or Fisher exact test.

Discussion

Principal Findings

This study shows that community pharmacies in Selangor are partly utilizing ICT apps/services. The Malaysian urban landscape is changing with the increased use of ICT-driven integrated smart cities, where businesses and services are making use of ICT apps for effective delivery and functioning [2,10]. However, Malaysian community pharmacies are adopting ICT relatively slowly. This study shows that young community pharmacists who are computer savvy and who are active internet users are leading these changes, with a greater proportion of younger than older respondents reporting using mobile or online apps for their online pharmacy stores and adopting labeling machines (Table 3).

Most of the community pharmacies in this study are located in residential areas, towns, and markets, and provide greater access to medicines to people living in and around these areas. This is a different approach to community pharmacy positioning because earlier community pharmacies were mostly set up in a nearby colony near hospitals or medical clinics to provide easier access to consumers [12]. As people moved from traditional colonies of houses to modern condominiums with integrated housing and other facilities, community pharmacies needed to

position themselves to serve their consumers better. However, the downside of the current positioning of community pharmacies might be that rural areas with sparse housing density may not have easier access to pharmacies. More rural clinics with pharmacies may have to be set up to bridge this gap of community pharmacies in rural areas [13].

Another aspect of ICT utilization is providing people with online access. Only some community pharmacies in our study had an online store, a mobile app, and products available via an online business platform such as Lazada and Shopee [14]. Unlike pharmacies, online trading platforms such as Lazada and Shopee are available for most of the other commercial products in Malaysia. Online pharmacies provide people with improved access, and they might provide them with cheaper medicines and the ability to save time and money, as consumers do not have to visit the community pharmacy in person [15]. However, regulating products in online pharmacies, especially health products such as cosmetics, nutraceuticals, and supplements imported from overseas, is a challenge, as these products might have different quality standards when they are manufactured in the host country. It is the responsibility of the regulating authorities to monitor online pharmacies, and the sales and distribution of medicines and health products via online business platforms [15,16]. In addition, privacy and security could be another concern with some forms of online purchasing, and there is some evidence suggesting that after the completion of consumers' details and monetary transactions, the consumers could ultimately not receive the products they ordered, posing problems such as frustration and delay of care [16].

ICT can help community pharmacies integrate with the digital ecosystems of urban cities. This study showed the ability to find the location of community pharmacies via GPS and other smart navigation systems. This provision will help consumers, as well as visiting tourists and others, access community pharmacies more easily, especially in cities where people use internet-based apps and navigation systems to access businesses and services. A study by Watson et al [17] on the health-seeking behavior of consumers showed that convenience of location affects patients' visits to community pharmacies for consultation for minor illnesses. The use of an ICT-based navigation system for locating pharmacies and the ability to check the availability of the necessary medicines will provide convenience as well as improved access to pharmacies for consumers.

ICT is an important means of communication and advertisement for different businesses and services [18]. This study showed that community pharmacies were using social media such as Facebook for advertisements, especially to improve the visibility and branding of pharmacy stores. A global study conducted by Benetoli et al [19] on participants from 8 countries (Australia, New Zealand, the United States, Brazil, Germany, Nigeria, Thailand, Philippines, and the United Kingdom) showed that social media positively impacted the pharmacy business by allowing community pharmacists to stay connected with customers or driving customers to their pharmacies [19]. Apart from advertisements, social media can be used for education and information dissemination, recruitment drives, and pharmacy services [19]. However, the dominant use of telephone, along with email, WhatsApp, and text messaging, shows that community pharmacies are diversifying the way they communicate and reach out to customers and patients. Staessen [20] reported that the use of reminders and follow ups with patients from community pharmacy staff via smartphone or text messaging services was effective in improving patients' adherence to medicines.

Automation is a key development that is driving change in the community pharmacy sector globally [21]. This study shows the minimum use of automated devices in pharmacies, such as pill-counting machines, automatic fillers, and prescription scanners. However, few pharmacies used label-printing machines, and very few had pill-counting machines and unit-dose packaging machines. Nevertheless, there was widespread use of ICT such as software for logistics and accounting purposes by community pharmacies in this study. Studies have shown that the use of modern ICT apps such as electronic inventory systems and barcode technology leads to improved technical accuracy and a reduction in the incidence of potential medication dispensing errors [22,23]. Carmenates and Keith [24] showed that automation in community pharmacies will lead to time savings, thereby providing more time for pharmacists to focus on delivering pharmaceutical care to the patient. Moreover, the present study shows the widespread use of electronic payment systems, including credit cards and new methods of payment such as Alipay, Epay, and Boost by community pharmacies. This shows that Malaysian community pharmacies are adopting ICT apps for business and transactions, and are slowly preparing themselves for the next technological revolution.

Approximately half of the community pharmacies included in this study used electronic patient records. The electronic record system allows pharmacies to record detailed information about the patients. For instance, in Australia, community pharmacies utilized ICT tools such as My Health Record to keep data of patients in cloud storage systems to ensure better medication management [25]. The use of an electronic record system saves the pharmacists time in going through the hard files of patient medical records and makes it easier for community pharmacies to communicate with the general practitioner.

Our study shows that community pharmacies also used mobile apps to gain access to medical databases. Apidi et al [26] reported that Lexicomp, Epocrates, Micromedex, and Drugs.com were the most commonly used drug information sources. Registered pharmacists in the United Kingdom reported using mobile apps to access drug databases such as the British National Formulary and Martindale because of the simplicity, user friendliness, and up-to-date information offered [27]. However, the study also reported the role of factors such as risks, company policies, and lack of regulations that may hinder the use of mobile health apps in community pharmacies. Few community pharmacies in Malaysia have developed their own mobile apps to promote their products and services. The personal details of their consumers are recorded as members, and these members can enjoy the privilege of discounted prices. According to a study carried out by DiDonato et al [28], several factors prompt consumers to use mobile health apps, such as easy accessibility, privacy assurance, and beneficiaries, and consumers were less likely to use mobile apps when there were

issues related to reliability, cost, and privacy. Mobile apps can help both the consumer/patient and community pharmacists. The greater challenge is to use mobile apps for patient self-management of diseases and to link the data from these mobile apps to pharmacies and health systems while creating an ecosystem that can regulate the use of these mobile health apps.

Strengths and Limitations

This study provides useful data on the ICT and electronic health infrastructures in community pharmacies in the state of Selangor, Malaysia. However, some limitations of this study should be acknowledged. First, the study was performed with only 60 community pharmacies, as some declined to participate due to time constraints, especially community pharmacies with limited staff (ie, 1 or 2 staff members). Consequently, the study might not be representative of all pharmacies in Selangor and in Malaysia, which could affect the generalizability of the results, especially in rural areas. However, given the limited literature from Malaysia on this topic, we believe that the study findings are helpful in providing future guidance.

Implications and Recommendations for Future Practice

We believe that more investment in ICT in the community pharmacy sector is needed in Malaysia. This approach would be aided by the good internet coverage in the state of Selangor and in Malaysia in general [29]. This could lead to several benefits to this sector. In particular, this could increase the business of community pharmacies, as clients and patients could have online access to all services and products, including online orders and delivery. It could also improve the population's quality of life by freeing the pharmacists' time to provide patient-centered services and counseling rather than focusing on time-consuming tasks than can be done perfectly by the technology. Additionally, it can help in reducing medication errors and other issues related to medication use. Furthermore, it can improve logistics and inventory management along with other managerial aspects of operating a pharmacy.

Conclusion

Community pharmacies in Malaysia are using ICT apps and adapting to the pharmacy services needs of modern smart cities. However, the adoption of ICT apps for pharmacy services has been slow and varied. The use of ICT apps for accounting, logistics, and similar tasks was high, whereas the use of ICT for automation, dispensing, and pharmaceutical care service delivery was relatively low. Future community pharmacies will be driven by patient-centered services and the use of ICT apps. Further studies in Malaysia, with country-wide coverage, need to focus on ICT usage in community pharmacies, consumer preferences, and regulatory ecosystems that guide ICT usage in pharmacies.

Acknowledgments

The authors would like to thank the School of Pharmacy, Monash University Malaysia, for undergraduate project support.

Authors' Contributions

BK conceptualized the study. LC, DL, and CW performed the data collection and wrote the first draft of the manuscript. BK, AB, JD, and AA revised and finalized the manuscript.

Conflicts of Interest

None declared.

References

1. Aceto G, Persico V, Pescapé A. The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges. J Netw Comput Appl 2018 Apr;107(1):125-154. [doi: 10.1016/j.jnca.2018.02.008]
2. Abas M. ICT Usage in Malaysia: A study on its economic impact dissertation on the internet. Tokyo: University of Waseda; 2005. URL: http://gits-db.jp/bulletin/2005/2005papers/2005dissertation_02_mohd.pdf [accessed 2019-01-25]
3. Enhancing the Contribution of Digitalisation to the Smart Cities of the Future. Organisation for Economic Co-operation and Development (OECD). Paris; 2019. URL: http://www.oecd.org/cfe/regional-policy/Smart-Cities-FINAL.pdf [accessed 2019-04-24]
4. Bigirimana S, Chinembiri M. Towards e-pharmacy: The future information and communication technologies needs for community pharmacies in Harare, Zimbabwe. Int J Econ Comm Manage 2015 Apr;3(4):1-26 [FREE Full text]
5. Leung V, Tharmalingam S, Cooper J, Charlebois M. Canadian community pharmacists' use of digital health technologies in practice. Can Pharm J 2016 Jan;149(1):38-45 [FREE Full text] [doi: 10.1177/1715163515618679] [Medline: 26798376]
6. Petrakaki D, Cornford T, Hibberd R, Lichtner V, Barber N. The role of technology in shaping the professional future of community pharmacists: The case of the electronic prescription service in the English National Health Service Internet. In: Chiasson M, Henfridsson O, Karsten H, DeGross J, editors. Researching the Future in Information Systems. IFIP Advances in Information and Communication Technology. Berlin: Springer; 2011:179-195.
7. Quek D. The Malaysian healthcare system: a review. 2009 Apr Presented at: Intensive workshop on health systems in transition April; 2009; Kuala Lumpur.

8. World Health Organization. The role of the pharmacist in the health care system: preparing the future pharmacist. Report of a third WHO Consultative Group on the Role of the Pharmacist, Vancouver, Canada, 27-29 August 1997. Geneva: World Health Organization; 1997. URL: https://apps.who.int/iris/bitstream/handle/10665/59169/WHO_PHARM_94.569.pdf?sequence=1&isAllowed=y [accessed 2019-08-22]
9. Howard P. The role of pharmacists. In: Pulcini C, Can F, Ergönül O, Beović B, editors. Antimicrobial Stewardship. Cambridge: Cambridge Academic Press; 2017:129-137.
10. Kassim NS. Smart city initiatives in Malaysia. PLAN Malaysia. 2011. URL: https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/Sesi-5-Rangkakerja-Berteraskan-Aspek-Informasi-(PLANMalaysia).pdf [accessed 2020-04-20]
11. Ranta P. Information and communications technology in health care. Economics Master's thesis. Helsinki: Aalto University; 2010. URL: http://epub.lib.aalto.fi/en/ethesis/pdf/12398/hse_ethesis_12398.pdf [accessed 2019-10-10]
12. Mak V, Hassali MAA. Separation of dispensing and prescribing in Malaysia: will the time come? J Pharm Pract Res 2015 Dec 29;45(4):394-395. [doi: 10.1002/jppr.1162]
13. Ariff K, Teng CL. Rural health care in Malaysia. Aust J Rural Health 2002 Apr;10(2):99-103. [doi: 10.1046/j.1440-1584.2002.00456.x] [Medline: 12047504]
14. Yee AYC, Kee DMH, Xing C, Qian PY, Qi SM, Dehrab AT. Lazada Group. In: J Int Conf Proc. 2019 Oct 21 Presented at: 4th International Conference on Project Management (ICPM); October 21, 2019; Manado p. 19-29. [doi: 10.32535/jicp.v2i2.599]
15. Fung CH, Woo HE, Asch SM. Controversies and legal issues of prescribing and dispensing medications using the Internet. Mayo Clin Proc 2004 Feb;79(2):188-194. [doi: 10.4065/79.2.188] [Medline: 14959914]
16. Orizio G, Merla A, Schulz PJ, Gelatti U. Quality of online pharmacies and websites selling prescription drugs: a systematic review. J Med Internet Res 2011 Sep 30;13(3):e74 [FREE Full text] [doi: 10.2196/jmir.1795] [Medline: 21965220]
17. Watson MC, Ferguson J, Barton GR, Maskrey V, Blyth A, Paudyal V, et al. A cohort study of influences, health outcomes and costs of patients' health-seeking behaviour for minor ailments from primary and emergency care settings. BMJ Open 2015 Feb 18;5(2):e006261 [FREE Full text] [doi: 10.1136/bmjopen-2014-006261] [Medline: 25694456]
18. Barrett J. Community pharmacies actively use social media, survey says. Pharmacy Times. New Jersey: Pharmacy & Healthcare Communications LLC; 2016 Sep 14. URL: https://www.pharmacytimes.com/news/community-pharmacies-actively-use-social-media-survey-says [accessed 2019-10-20]
19. Benetoli A, Chen TF, Schaefer M, Chaar BB, Aslani P. Professional Use of Social Media by Pharmacists: A Qualitative Study. J Med Internet Res 2016 Sep 23;18(9):e258 [FREE Full text] [doi: 10.2196/jmir.5702] [Medline: 27663570]
20. Staessen J. Technology to improve adherence in community pharmacy: a literature review. J Pharm Belg 2015 Mar(1):16-23. [Medline: 26571793]
21. Angelo LB, Christensen DB, Ferreri SP. Impact of community pharmacy automation on workflow, workload, and patient interaction. J Am Pharm Assoc 2005 Mar;45(2):138-144. [doi: 10.1331/1544345053623537] [Medline: 15868755]
22. Oldland A, Golightly L, May S, Barber G, Stolpman N. Electronic Inventory Systems and Barcode Technology: Impact on Pharmacy Technical Accuracy and Error Liability. Hospital Pharm 2015 Jan;50(1):034-041. [doi: 10.1310/hpj5001-034]
23. Flynn EA, Barker KN. Effect of an automated dispensing system on errors in two pharmacies. J Am Pharm Assoc 2006 Sep;46(5):613-615. [doi: 10.1331/1544-3191.46.5.613.flynn] [Medline: 17036648]
24. Carmenates J, Keith MR. Impact of automation on pharmacist interventions and medication errors in a correctional health care system. Am J Health Syst Pharm 2001 May 01;58(9):779-783. [doi: 10.1093/ajhp/58.9.779] [Medline: 11351917]
25. What's Trending: Technology in Pharmacy. Austral J Pharm. New South Wales; 2017. URL: https://ajp.com.au/features/whats-trending-technology-pharmacy/ [accessed 2019-01-25]
26. Apidi NA, Murugiah MK, Muthuveloo R, Soh YC, Caruso V, Patel R, et al. Mobile Medical Applications for Dosage Recommendation, Drug Adverse Reaction, and Drug Interaction: Review and Comparison. Ther Innov Regul Sci 2017 Jul;51(4):480-485. [doi: 10.1177/2168479017696266] [Medline: 30227053]
27. Davies MJ, Collings M, Fletcher W, Mujtaba H. Pharmacy Apps: a new frontier on the digital landscape? Pharm Pract 2014 Jul;12(3):453 [FREE Full text] [doi: 10.4321/s1886-36552014000300009] [Medline: 25243034]
28. DiDonato KL, Liu Y, Lindsey CC, Hartwig DM, Stoner SC. Community pharmacy patient perceptions of a pharmacy-initiated mobile technology app to improve adherence. Int J Pharm Pract 2015 Oct;23(5):309-319. [doi: 10.1111/ijpp.12168] [Medline: 25572628]
29. Accelerating the rise of a smart nation (2017 annual report). Malaysian Communications and Multimedia Commission (MCMC). 2017. URL: https://www.skmm.gov.my/skmmgovmy/media/General/pdf/AR-2017-Eng.pdf [accessed 2020-04-20]

Abbreviations

ICT: information and communication technology

Edited by G Eysenbach; submitted 27.01.20; peer-reviewed by S Khanal, J Gauly; comments to author 29.02.20; revised version received 11.05.20; accepted 13.05.20; published 08.07.20.

Please cite as:
KC B, Lim D, Low CC, Chew C, Blebil AQ, Dujaili JA, Alrasheedy AA
Positioning and Utilization of Information and Communication Technology in Community Pharmacies of Selangor, Malaysia: Cross-Sectional Study
JMIR Med Inform 2020;8(7):e17982
URL: https://medinform.jmir.org/2020/7/e17982
doi:10.2196/17982
PMID:32463787

Original Paper

Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods: Content Analysis

Xiaofeng Wang[1*], PhD; Shuai Chen[2*], MS; Tao Li[2], MS; Wanting Li[2], BS; Yejie Zhou[2], BS; Jie Zheng[2], BS; Qingcai Chen[2], PhD; Jun Yan[3], PhD; Buzhou Tang[2], PhD

[1]School of Communication, Shenzhen University, Shenzhen, China

[2]Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China

[3]Yidu Cloud (Beijing) Technology Co Ltd, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Buzhou Tang, PhD
Department of Computer Science
Harbin Institute of Technology Shenzhen Graduate School
L1407
Shenzhen
China
Phone: 86 13725525983
Email: tangbuzhou@gmail.com

Abstract

Background: Depression is a serious personal and public mental health problem. Self-reporting is the main method used to diagnose depression and to determine the severity of depression. However, it is not easy to discover patients with depression owing to feelings of shame in disclosing or discussing their mental health conditions with others. Moreover, self-reporting is time-consuming, and usually leads to missing a certain number of cases. Therefore, automatic discovery of patients with depression from other sources such as social media has been attracting increasing attention. Social media, as one of the most important daily communication systems, connects large quantities of people, including individuals with depression, and provides a channel to discover patients with depression. In this study, we investigated deep-learning methods for depression risk prediction using data from Chinese microblogs, which have potential to discover more patients with depression and to trace their mental health conditions.

Objective: The aim of this study was to explore the potential of state-of-the-art deep-learning methods on depression risk prediction from Chinese microblogs.

Methods: Deep-learning methods with pretrained language representation models, including bidirectional encoder representations from transformers (BERT), robustly optimized BERT pretraining approach (RoBERTa), and generalized autoregressive pretraining for language understanding (XLNET), were investigated for depression risk prediction, and were compared with previous methods on a manually annotated benchmark dataset. Depression risk was assessed at four levels from 0 to 3, where 0, 1, 2, and 3 denote no inclination, and mild, moderate, and severe depression risk, respectively. The dataset was collected from the Chinese microblog Weibo. We also compared different deep-learning methods with pretrained language representation models in two settings: (1) publicly released pretrained language representation models, and (2) language representation models further pretrained on a large-scale unlabeled dataset collected from Weibo. Precision, recall, and F1 scores were used as performance evaluation measures.

Results: Among the three deep-learning methods, BERT achieved the best performance with a microaveraged F1 score of 0.856. RoBERTa achieved the best performance with a macroaveraged F1 score of 0.424 on depression risk at levels 1, 2, and 3, which represents a new benchmark result on the dataset. The further pretrained language representation models demonstrated improvement over publicly released prediction models.

Conclusions: We applied deep-learning methods with pretrained language representation models to automatically predict depression risk using data from Chinese microblogs. The experimental results showed that the deep-learning methods performed better than previous methods, and have greater potential to discover patients with depression and to trace their mental health conditions.

(JMIR Med Inform 2020;8(7):e17958) doi:10.2196/17958

KEYWORDS

depression risk prediction; deep learning; pretrained language model; Chinese microblogs

Introduction

Background

Mental health is an important component of personal well-being and public health as reported by the World Health Organization (WHO) [1]. Anyone—regardless of gender, financial status, and age—may suffer from mental disorders, among which depression remains the most common form [2]. Depression is reported to affect more than 264 million people worldwide according to the WHO's Comprehensive Mental Health Action Plan 2003-2020 [3], and the number has been quickly increasing in recent years [4]. Among various depressive illnesses, the lifetime prevalence of major depressive disorders is approximately 16%, and evidence suggests that the incidence is increasing [5]. In 1997, the WHO estimated that depression will be the second most debilitating disease by 2020, behind cardiovascular disease [6].

Depression is accompanied by a suite of very negative effects, as it can interfere with a person's daily life and routine. In the short term, depression may reduce an individual's enjoyment of life, make them withdraw from their family and friends, and ultimately feel lonely. In the long term, prolonged depression may lead to more serious conditions and illnesses. Fortunately, early recognition and treatment are proven to be helpful for people with depression to reduce the negative impacts of the disorder [7]. Despite broad developments in medical technology, it remains difficult to diagnose depression due to the particularity of mental disorders [8]. Currently, most diagnoses of depressive illness are based on self-reports or self-diagnosis of patients [9,10]. The diagnosis procedures are complex and time-consuming. Moreover, a high proportion of patients with depression cannot be discovered as they do not want to disclose or discuss their mental health conditions with others. Therefore, it is urgent to find methods that can help to discover patients with depression from other channels.

With the development of information technology, social media has become an important part of people's daily life. More and more people are using social media platforms such as Twitter, Facebook, and Sina Weibo to share their thoughts, feelings, and emotional status. These social media platforms can provide a huge amount of valuable data for research. Some studies based on social media data such as personalized news recommendation [11], public opinion sensing and trend analysis [12], disease transmission trend monitoring [13], and future patient visits prediction [14] have achieved good results. In the case of depression, as social media platforms have become important forums for people with depression to interact with peers within a comfortable emotional distance [15], high numbers of patients with depression tend to gather to share their feelings, emotional status, and treatment procedures. Some researchers have attempted to discover patients with depression from social media, such as by predicting depression risk embedded in text from microblogs. Accumulating evidence shows that the language and emotion posted on social media platforms could indicate depression [3].

In this study, we investigated the use of deep-learning methods for depression risk prediction from data collected in Chinese microblogs. This study represents an extension of the study of Wang et al [16], who presented an annotated dataset of Chinese microblogs for depression risk prediction and compared four machine-learning methods, including the deep-learning method bidirectional encoder representations from transformers (BERT) [17]. Here, we further investigated three deep-learning methods with pretrained language representation models, BERT, robustly optimized BERT pretraining approach (RoBERTa) [18], and generalized autoregressive pretraining for language understanding (XLNET) [19], on the depression dataset and obtained new benchmark results.

Related Work

In early studies focused on depression detection, most of the methods applied were rule-based and those based on self-reporting or self-diagnosis. For example, Hamilton [20] established a rating scale for depression to help patients with depression evaluate the severity of their depression by themselves according to a self-report. However, these methods always require domain experts to define the rules and are time-consuming. In recent years, with the rapid spread of social media, more and more information about personal daily life is publicly posted on the internet, which can be widely used for health prediction, including depression detection.

Choudhury et al [9] made a major contribution to the field of depression detection from social media by investigating whether social media can be used as a source of information to detect mental illness among individuals as well as within a population. Following this study, several researchers annotated some corpora for automatic depression detection, including depression level prediction. For example, Glen et al [21] constructed an annotated corpus composed of 1746 users collected from Twitter for depression detection. In the corpus, the users were divided into three groups: depression users, posttraumatic stress disorder (PTSD) users, and control users. This corpus was used as the dataset of the Computational Linguistics and Clinical Psychology (CLPsych) shared task in 2015 [22] to predict PTSD users from the control group, users with depression from the control group, and users with depression among users with PTSD. The system that ranked first in the CLPsych 2015 shared task was a combination system composed of 16 support vector machine (SVM)-based subsystems based on features derived using supervised linear discriminant analysis [23], supervised Anchor (for topic modeling), and lexical term frequency-inverse document frequency [24]. Cacheda et al [25] presented a social network analysis and random forest algorithm to detect early depression. Ricard et al [26] trained an elastic-net regularized linear regression model on Instagram post captions and comments to detect depression. The features used in the linear

regression model included multiple sentiment scores, emoji sentiment analysis results, and metavariables such as the number of "likes" and average comment length. Lin et al [27] proposed a deep neural network model to detect users' psychological stress by incorporating two different types of user-scope attributes, and evaluated the model on four different datasets from major microblog platforms, including Sina Weibo, Tencent Weibo, and Twitter. Most of these studies focused on user-level depression detection, as summarized by Wongkoblap et al [28], and the machine-learning methods used in these studies included SVM, logistic regression, decision trees [29-32], random forest [33,34], naive Bayes [35,36], K-nearest neighbor, maximum entropy [37], neural network, and deep-learning neural network.

To analyze social media at a fine-granularity level and track the mental health conditions of patients with depression, some researchers attempted to detect depression at the tweet level. Jamil et al [38] constructed two types of datasets from Twitter for depression detection: one annotated at the tweet level consisting of 8753 tweets and the other annotated at the user level consisting of 160 users. The SVM-based system developed on these two datasets performed well at the user level, but not very well at the tweet level. Wang et al [16] annotated a dataset from Sina Weibo at the microblog level (equivalent to the tweet level), in which each microblog was labeled with a depression risk ranging from 0 to 3. They compared four machine-learning methods on this dataset, including SVM, convolutional neural network (CNN), long short-term memory network (LSTM), and BERT. The three deep-learning methods (ie, CNN, LSTM, and BERT) significantly outperformed SVM, and BERT showed the best performance among them.

During the last 2 or 3 years, pretrained language representation models such as BERT, RoBERTa, and XLNET have shown significant performance gains in many natural language processing tasks such as text classification, question answering, and others [39]. However, to the best of our knowledge, deep-learning methods with pretrained language representation models have not yet been applied to depression risk prediction.

Methods

Dataset

In this study, we use the dataset provided by Wang et al [16], which was collected from the Chinese social media platform Sina Weibo. In this dataset, 13,993 microblogs were annotated with depression risk assessed at four levels from 0 to 3, where 0 indicates no inclination to depression, or only some common pressures such as work, study, and family issues; 1 indicates mild depression, denoting that users express despair with life but do not mention suicide or self-harm; 2 indicates moderate depression, which denotes that users mention suicide or self-harm without stating a specific time or place; and 3 indicates severe depression, which denotes that users mention suicide or self-harm with a specific time or place. A total of 11,835 microblogs were annotated as 0, 1379 microblogs were annotated as 1, 650 microblogs were annotated as 2, and the remaining 129 microblogs were annotated as 3. The distribution of microblogs at different levels was imbalanced. Table 1 provides examples of the different depression levels. Following Wang et al [16], we split the dataset into two parts: a training set of 11,194 microblogs and a test set of 2799 microblogs, as shown in Table 2.

Table 1. Examples of different depression risk levels in the dataset.

Depression risk level	Microblog
3	Weibo: 不出意外的话，我打算死在今年 。 Barring accidents, I plan to commit suicide this year.
2	Weibo: 我一直策划着如何自杀，可是放不下的太多了。 I have been planning to commit suicide, but I cannot let go of too many things.
1	Weibo: 如果我累，真的离开了。 If I'm tired, I will leave.
0	Weibo: 吃了个早餐应该能维持今天。 The breakfast I ate should be able to support me today.

Table 2. Dataset statistics.

Depression level	Training set (n)	Test set (n)
3	103	26
2	520	130
1	1103	276
0	9468	2367
All	11,194	2799

Deep-Learning Methods Based on Pretrained Language Representation Models

BERT

BERT is a language representation model designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both the left and right context in all layers [17]. It uses the transformer architecture to capture long-distance dependences in sentences. During pretraining, BERT optimizes the masked language model (MLM) and the next sentence prediction (NSP) task jointly on large-scale unlabeled text. To implement NSP, BERT adds the token [CLS] at the beginning of every sequence. The final hidden state corresponding to the token [CLS] is then used as the aggregate sequence representation for downstream tasks. When the language representation model is pretrained, it can be subsequently fine-tuned for downstream tasks using the labeled data of downstream tasks. BERT achieved better performance on several natural language processing tasks in 2018 [17]. In the present study, depression risk prediction was formalized as a classification task; therefore, we simply needed to feed the representation of token [CLS] into an output layer (a fully connected layer) and then fine-tune the whole network.

RoBERTa

RoBERTa is an optimized replication version of BERT [18]. Compared with BERT, RoBERTa offers the following four improvements during training: (1) training the model for a longer period with larger batches over more data; (2) removing the NSP task; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data. Based on these improvements, RoBERTa has achieved new state-of-the-art results on many tasks compared with BERT [18].

XLNET

XLNET is a generalized autoregressive method that takes advantage of both autoregressive language modeling and autoencoding while avoiding their limitations [19]. As BERT and its variants (eg, RoBERTa) neglect the dependency between the masked positions and suffer from a pretrain-finetune discrepancy, XLNET adopts a permutation language model instead of MLM to solve the discrepancy problem. For downstream tasks, the fine-tuning procedure of XLNET is similar to that of BERT and RoBERTa.

Experiments

Experimental Setup

We investigated the different deep-learning methods with pretrained language representation models in two settings: (1) publicly released pretrained language representation models and (2) language representation models further pretrained on a large-scale unlabeled dataset collected from Weibo based on (1). The hyperparameters for BERT, RoBERTa, and XLNET for depression risk prediction are listed in Table 3. These hyperparameters were obtained by crossvalidation.

Table 3. Hyperparameters for the deep-learning methods.

Parameter	BERT[a]	RoBERTa[b]	XLNET[c]
Learning rate	1e-5	1e-5	2e-5
Training steps	7000	7000	7000
Maximum length	128	128	128
Batch size	16	16	16
Warm-up steps	700	700	700
Dropout rate	0.3	0.3	0.3

[a]BERT: bidirectional encoder representations from transformers.

[b]RoBERTa: robustly optimized bidirectional encoder representations from transformers pretraining approach.

[c]XLNET: generalized autoregressive pretraining for language understanding.

In-Domain Pretraining

For in-domain pretraining (IDP), we started from the public released pretrained BERT model [40], RoBERTa model [41], and XLNET model [42], and further pretrained them on the same unlabeled Weibo corpus as used by Wang et al [16]. The unlabeled corpus contains about 300,000 microblogs. The hyperparameters used during further IDP are listed in Table 4. These hyperparameters were optimized by crossvalidation.

Table 4. Hyperparameters during further in-domain pretraining for the deep-learning methods.

Parameter	BERT[a]	RoBERTa[b]	XLNET[c]
Learning rate	2e-5	2e-5	2e-5
Training steps	100,000	100,000	100,000
Maximum length	256	256	256
Batch size	16	16	16
Warm-up steps	10,000	10,000	10,000

[a]BERT: bidirectional encoder representations from transformers.

[b]RoBERTa: robustly optimized bidirectional encoder representations from transformers pretraining approach.

[c]XLNET: generalized autoregressive pretraining for language understanding.

Evaluation Criteria

Micro/macro precision, recall, and the F1 score were used to evaluate the performance of the different deep-learning methods.

Results

Table 5 shows the performance of deep-learning methods with different language representation models. For each deep-learning method, the addition of a pretrained language representation model brought improvement over the publicly released language representation model. Among the three methods, BERT showed the best performance, with the highest microF1 score of 0.856 (BERT_IDP). The microF1 score difference between any two of the three methods was around 1%-2%, which is not satisfactory. Compared with CNN and LSTM, BERT, RoBERTa, and XLNET showed a great advantage.

Almost all of the deep-learning methods performed the best on level 0 and performed the worst on level 3, which may be caused by data imbalance. For all depression risk levels except for level 0, the deep-learning methods showed different performance rankings. On level 1, RoBERTa_IDP performed the best with an F1 score of 0.422, whereas on level 2, XLNET_IDP achieved the best F1 score of 0.493, and on level 3, XLNET achieved the best F1 score of 0.445.

As the aim of this study was to discover potential patients with depression, we were more interested in microblogs at levels 1, 2, and 3. Therefore, it is more meaningful to report macro precision, recall, and F1 scores on these three levels, which are shown in Table 6, in which the highest values in each column are in italics. The advantage of RoBERTa_IDP for microblog-level depression detection can be clearly seen. The confusion matrices of BERT_IDP, RoBERTa_IDP, and XLNET_IDP are shown in Table 7.

Table 5. Performance of deep-learning methods with different language representation models.

Model	Level-0			Level-1			Level-2			Level-3			MicroF1
	P[a]	R[b]	F1	P	R	F1	P	R	F1	P	R	F1	
CNN[c] [16]	0.908	0.940	0.924	0.380	0.236	0.291	0.351	0.415	0.380	0.250	0.231	0.240	0.841
LSTM[d] [16]	0.896	0.936	0.916	0.294	0.288	0.257	0.324	0.262	0.289	0.714	0.192	0.303	0.832
BERT[e] [16]	0.942	0.894	0.917	0.323	0.502	0.393	0.468	0.489	0.478	0.574	0.152	0.240	0.834
BERT_IDP[f] [16]	0.929	0.938	*0.934* [g]	0.394	0.446	0.418	0.568	0.385	0.459	0.667	0.231	0.343	*0.856*
RoBERTa[h]	0.931	0.920	0.925	0.355	0.464	0.402	0.556	0.385	0.455	0.600	0.231	0.333	0.843
RoBERTa_IDP	0.933	0.920	0.926	0.371	0.489	*0.422*	0.578	0.400	0.473	0.636	0.269	0.333	0.847
XLNET[i]	0.908	0.948	0.927	0.358	0.273	0.309	0.484	0.353	0.408	0.530	0.384	*0.445*	0.848
XLNET_IDP	0.933	0.920	0.926	0.361	0.471	0.409	0.577	0.431	*0.493*	0.625	0.192	0.294	0.846

[a]P: precision.

[b]R: recall.

[c]CNN: convolutional neural network.

[d]LSTM: long short-term memory network.

[e]BERT: bidirectional encoder representations from transformers.

[f]_IDP: The model is further trained on the in-domain unlabeled corpus.

[g]Highest F1 values are indicated in italics.

[h]RoBERTa: robustly optimized bidirectional encoder representations from transformers pretraining approach.

[i]XLNET: generalized autoregressive pretraining for language understanding.

Table 6. Performance of deep-learning methods with different language representation models on level 1, 2 and 3.

Model	Macro-F1	Macro-P[a]	Macro-R[b]
BERT[c] [16]	0.370	0.455	0.381
BERT_IDP[d] [16]	0.406	*0.543* [e]	0.354
RoBERTa[f]	0.396	0.503	0.360
RoBERTa_IDP	*0.424*	0.528	*0.386*
XLNET[g]	0.387	0.457	0.336
XLNET_IDP	0.398	0.521	0.364

[a]P: precision.

[b]R: recall.

[c]BERT: bidirectional encoder representations from transformers.

[d]_IDP: The model is further trained on the in-domain unlabeled corpus.

[e]Highest F1 values are indicated in italics.

[f]RoBERTa: robustly optimized bidirectional encoder representations from transformers pretraining approach.

[g]XLNET: generalized autoregressive pretraining for language understanding.

Table 7. Confusion matrix of the deep-learning methods with in-domain training.

Gold-standard method	Prediction method Level-0	Prediction method Level-1	Prediction method Level-2	Prediction method Level-3
BERT_IDP[a]				
Level-0	2221	131	14	1
Level-1	137	123	16	0
Level-2	26	52	50	2
Level-3	6	6	8	6
RoBERTa_IDP[b]				
Level-0	2177	176	13	1
Level-1	128	135	15	0
Level-2	26	47	52	3
Level-3	3	6	10	7
XLNET_IDP[c]				
Level-0	2177	176	13	1
Level-1	128	130	18	0
Level-2	26	46	56	2
Level-3	3	8	10	5

[e]BERT_IDP: bidirectional encoder representations from transformers further trained on the in-domain unlabeled corpus.

[b]RoBERTa_IDP: robustly optimized bidirectional encoder representations from transformers pretraining approach further trained on the in-domain unlabeled corpus.

[c]XLNET_IDP: generalized autoregressive pretraining for language understanding further trained on the in-domain unlabeled corpus.

Discussion

Principal Findings

In this study, we have applied three deep-learning methods with pretrained language representation models to predict the depression risk based on data from Chinese microblogs, which is recognized as a text classification task. The deep-learning methods achieved the highest macroaveraged F1 score of 0.424 on the three levels of depression of concern, which represents a new state-of-the-art result from the dataset used by Wang et al [16]. These results indicate the potential for tracing mental health conditions of depression patients from microblogs. We also investigated the effect of pretraining language representation models in different settings. These experiments showed that further applying pretrained language representation models on a large-scale unlabeled in-domain corpus leads to better performance, which is easily interpretable.

Error analysis on the deep-learning methods showed that several errors often occur between level 0 and level 1. As shown in the confusion matrix in Table 7, among all samples predicted incorrectly by RoBERTa_IDP, 128 gold-standard samples at level 1 were predicted as level 0 and 176 gold-standard samples at level 0 were predicted as level 1. This type of error accounted for about 70% of all errors. The main reason for this phenomenon is that there are many ambiguous words in Chinese microblogs, which are difficult to be distinguished independently. These ambiguous words also occurred very frequently in microblogs of high depression risk levels. For example, in microblog "我已经放下了亲情、友情，都已经和解了，可以安心上路了(I have let go of my family and friendships, and have reconciled with them. Now, I can go on my way with ease)," "上路" is an ambiguous word. In Chinese, this word not only means "going on one's way" but also has the meaning of passing away. Other examples include "解脱(extricate)" in "啥时候能够解脱呢？有点期待 (When can I extricate myself from the tough world? I am looking forward to it)," and "黑(black)" in "我看到的世界都是黑的只剩下一片黑 (The world I see is black, only black)." These words are not related to depression risk in most common contexts. However, in the contexts mentioned above, these words indicate the despair of patients in life. Since these words appeared infrequently in the entire depression dataset, it was very difficult for the deep-learning models to learn the multiple meanings of these ambiguous words. From the confusion matrix, we can see that RoBERTa_IDP could correctly classify more samples at a high level than the previous BERT model. This suggests that our new methods can handle these types of errors better than previous methods. For these types of errors, there may be two possible solutions: one is to import more samples containing these ambiguous words to help the models learn the multiple meanings of these words, and the other is to import more of the context from the same user to help the models make a correct prediction.

In the future, there may be three directions for further improvement. First, we will expand the current dataset to cover as many multiple meanings of ambiguous words as possible. Second, we will attempt to use user-level context to improve microblog-level depression risk prediction. Third, we will try to add medical knowledge regarding depression into the deep-learning methods.

Conclusion

Depression is one of the most harmful mental disorders worldwide. The diagnosis of depression is quite complex and time-consuming. Predicting depression risk automatically is very important and meaningful. In this study, we have focused on the potential of deep-learning methods with pretrained language representation models for depression risk prediction from Chinese microblogs. The experimental results on a benchmark dataset showed that the proposed methods performed well for this task. The main contribution of this study to depression health care is to help discover potential patients with depression from social media quickly. This could help doctors or psychologists to concentrate on providing help for these potential patients with a high depression level.

Acknowledgments

This study is supported in part by grants from the National Natural Science Foundations of China (U1813215, 61876052, and 61573118), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), National Natural Science Foundations of Guangdong, China (2019A1515011158), Guangdong Province Covid-19 Pandemic Control Research Fund (2020KZDZX1222), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20180306172232154 and JCYJ20170307150528934), and Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

Authors' Contributions

The work presented herein was carried out with collaboration among all authors. XW, SC, and BT designed the methods and experiments. XW and SC conducted the experiment. All authors analyzed the data and interpreted the results. SC and BT wrote the paper. All authors have approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Promoting mental health: Concepts, emerging evidence, practice: Summary report. World Health Organization. 2004. URL: https://www.who.int/mental_health/evidence/en/promoting_mhh.pdf [accessed 2020-07-07]
2. Results from the 2013 National Survey on Drug Use and Health: Mental Health Findings.: US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality; 2013. URL: https://www.samhsa.gov/data/sites/default/files/NSDUHmhfr2013/NSDUHmhfr2013.pdf [accessed 2020-07-07]
3. Saxena S, Funk M, Chisholm D. World Health Assembly adopts Comprehensive Mental Health Action Plan 2013-2020. Lancet 2013 Jun 08;381(9882):1970-1971 [FREE Full text] [doi: 10.1016/S0140-6736(13)61139-3] [Medline: 23746771]

4. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. Lancet 2007 Sep 08;370(9590):851-858. [doi: 10.1016/S0140-6736(07)61415-9] [Medline: 17826170]
5. Doris A, Ebmeier K, Shajahan P. Depressive illness. Lancet 1999 Oct;354(9187):1369-1375. [doi: 10.1016/s0140-6736(99)03121-9]
6. Murray CJ, Lopez AD. Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. Lancet 1997 May 17;349(9063):1436-1442. [doi: 10.1016/S0140-6736(96)07495-8] [Medline: 9164317]
7. Picardi A, Lega I, Tarsitani L, Caredda M, Matteucci G, Zerella M, SET-DEP Group. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. J Affect Disord 2016 Jul 01;198:96-101. [doi: 10.1016/j.jad.2016.03.025] [Medline: 27015158]
8. Baik S, Bowers BJ, Oakley LD, Susman JL. The recognition of depression: the primary care clinician's perspective. Ann Fam Med 2005 Jan 01;3(1):31-37 [FREE Full text] [doi: 10.1370/afm.239] [Medline: 15671188]
9. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. : Association for the Advancement of Artificial Intelligence; 2013 Jul 8 Presented at: Proceedings of the seventh international AAAI conference on weblogs and social media; 2013; Cambridge, MA, USA.
10. Sanchez-Villegas A, Schlatter J, Ortuno F, Lahortiga F, Pla J, Benito S, et al. Validity of a self-reported diagnosis of depression among participants in a cohort study using the Structured Clinical Interview for DSM-IV (SCID-I). BMC Psychiatry 2008 Jun 17;8(1):43. [doi: 10.1186/1471-244x-8-43]
11. Abel F, Houben GJ, Tao K. Analyzing user modeling on twitter for personalized news recommendations. In: Konstan JA, Conejo R, Marzo JL, Oliver N, editors. User Modeling, Adapatation and Personalization. UMAP 2011. Lecture Notes in Computer Science, vol. 6787. Berlin, Heidelberg: Springer; 2011.
12. Mingyi G, Renwei Z. A Research on Social Network Information Distribution Pattern With Internet Public Opinion Formation. Journalism Communication 2009;5:72-78.
13. Rothenberg RB, Sterk C, Toomey KE, Potterat JJ, Johnson D, Schrader M, et al. Using social network and ethnographic tools to evaluate syphilis transmission. Sex Transm Dis 1998 Mar;25(3):154-160. [doi: 10.1097/00007435-199803000-00009] [Medline: 9524994]
14. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of Predicting Health Care Utilization Via Web Search Behavior: A Data-Driven Analysis. J Med Internet Res 2016 Sep 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
15. Colineau N, Paris C. Talking about your health to strangers: understanding the use of online social networks by patients. New Rev Hypermedia Multimed 2010 Apr;16(1-2):141-160. [doi: 10.1080/13614568.2010.496131]
16. Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, et al. Assessing depression risk in Chinese microblogs: a corpus and machine learning methods. 2019 Presented at: IEEE International Conference on Healthcare Informatics (ICHI); June 10-13, 2019; Xi'an, China. [doi: 10.1109/ichi.2019.8904506]
17. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 2018:181004805.
18. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint 2019:1907.11692v1.
19. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint 2019:1906.08237.
20. Hamilton M. The Hamilton rating scale for depression. In: Sartorius N, Ban TA, editors. Assessment of depression. Berlin, Heidelberg: Springer-Verlag; 1986:143-152.
21. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. 2014 Presented at: Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality; June 2014; Baltimore, MD p. 51-60. [doi: 10.3115/v1/w14-3207]
22. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 shared task: Depression and PTSD on Twitter. 2015 Presented at: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; June 5, 2015; Denver, Colorado. [doi: 10.3115/v1/w15-1204]
23. Blei DM, Ng AY, Jordan MI. Latent dirichllocation. J Machine Learn Res 2003;3:993-1022.
24. Resnik P, Armstrong W, Claudino L, Nguyen T. The University of Maryland CLPsych 2015 shared task system. 2015 Jun 5 Presented at: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; June 5, 2015; Denver, Colorado. [doi: 10.3115/v1/w15-1207]
25. Cacheda F, Fernandez D, Novoa FJ, Carneiro V. Early Detection of Depression: Social Network Analysis and Random Forest Techniques. J Med Internet Res 2019 Jun 10;21(6):e12554 [FREE Full text] [doi: 10.2196/12554] [Medline: 31199323]
26. Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the Utility of Community-Generated Social Media Content for Detecting Depression: An Analytical Study on Instagram. J Med Internet Res 2018 Dec 06;20(12):e11817 [FREE Full text] [doi: 10.2196/11817] [Medline: 30522991]

27. Lin H, Jia J, Guo Q, Xue Y, Li Q, Huang J, et al. User-level psychological stress detection from social media using deep neural network. 2014 Nov 1 Presented at: Proceedings of the 22nd ACM international conference on Multimedia; November 2014; Orlando, FL. [doi: 10.1145/2647868.2654945]
28. Wongkoblap A, Vadillo MA, Curcin V. Researching Mental Health Disorders in the Era of Social Media: Systematic Review. J Med Internet Res 2017 Jun 29;19(6):e228 [FREE Full text] [doi: 10.2196/jmir.7215] [Medline: 28663166]
29. Burnap P, Colombo W, Scourfield J. Machine classification and analysis of suicide-related communication on twitter. 2015 Sep 1 Presented at: Proceedings of the 26th ACM Conference on Hypertext & Social Media; August 2015; Guzelyurt, Northern Cyprus p. 75-84. [doi: 10.1145/2700171.2791023]
30. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: a good place to detect health conditions. PLoS One 2014 Jan 29;9(1):e86191 [FREE Full text] [doi: 10.1371/journal.pone.0086191] [Medline: 24489699]
31. Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A depression detection model based on sentiment analysis in micro-blog social network. In: Li J, editor. Trends and Applications in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7867. Berlin, Heidelberg: Springer; Apr 14, 2013.
32. Wang X, Zhang C, Sun L. An improved model for depression detection in micro-blog social network. 2013 Dec 7 Presented at: IEEE 13th International Conference on Data Mining Workshops; December 7-10, 2013; Dallas, TX p. 2013. [doi: 10.1109/icdmw.2013.132]
33. Saravia E, Chang C, De LR, Chen YS. MIDAS: Mental illness detection and analysis via social media. 2016 Aug 18 Presented at: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); August 18-21, 2016; San Francisco, CA p. 2016. [doi: 10.1109/asonam.2016.7752434]
34. Guan L, Hao B, Cheng Q, Yip PS, Zhu T. Identifying Chinese Microblog Users With High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model. JMIR Ment Health 2015 May 12;2(2):e17 [FREE Full text] [doi: 10.2196/mental.4227] [Medline: 26543921]
35. Wang T, Brede M, Ianni A. Detecting and characterizing eating-disorder communities on social media. 2017 Feb 1 Presented at: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining; 2017; Cambridge, UK. [doi: 10.1145/3018661.3018706]
36. Hao B, Li L, Li A, Zhu T. Predicting mental health status on social media. In: Rau PLP, editor. Cross-cultural Design. Cultural Differences in Everyday Life. CCD 2013. Lecture Notes in Computer Science, vol 8024. Berlin, Heidelberg: Spring; Apr 23, 2014.
37. Mitchell M, Hollingshead K, Coppersmith G. Quantifying the language of schizophrenia in social media. 2015 Jan 1 Presented at: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; June 5, 2015; Denver, CO. [doi: 10.3115/v1/w15-1202]
38. Jamil Z, Inkpen D, Buddhitha P, White K. Monitoring tweets for depression to detect at-risk users. 2018 Aug 1 Presented at: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; August 2017; Vancouver, BC. [doi: 10.18653/v1/w17-3104]
39. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding with unsupervised learning. OpenAI. 2018 Jun 11. URL: https://openai.com/blog/language-unsupervised/ [accessed 2020-07-07]
40. bert. github. URL: https://github.com/google-research/bert [accessed 2020-07-07]
41. fairseq. github. URL: https://github.com/pytorch/fairseq [accessed 2020-07-07]
42. xlnet. github. URL: https://github.com/zihangdai/xlnet [accessed 2020-07-07]

Abbreviations

BERT: bidirectional encoder representations from transformers
CLPsych: Computational Linguistics and Clinical Psychology
CNN: convolutional neural network
IDP: in-domain pretraining
LSTM: long short-term memory network
MLM: masked language model
NSP: next sentence prediction
PTSD: posttraumatic stress disorder
RoBERTa: robustly optimized bidirectional encoder representations from transformers pretraining approach
SVM: support vector machine
WHO: World Health Organization
XLNET: generalized autoregressive pretraining for language understanding

Edited by J Bian; submitted 24.01.20; peer-reviewed by X Yang, L Zhang, G Lim; comments to author 04.04.20; revised version received 30.05.20; accepted 01.06.20; published 29.07.20.

Please cite as:
Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, Chen Q, Yan J, Tang B
Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods: Content Analysis
JMIR Med Inform 2020;8(7):e17958
URL: http://medinform.jmir.org/2020/7/e17958/
doi:10.2196/17958
PMID:32723719

Original Paper

Identifying and Predicting Intentional Self-Harm in Electronic Health Record Clinical Notes: Deep Learning Approach

Jihad S Obeid[1], MD; Jennifer Dahne[1], PhD; Sean Christensen[1], MD; Samuel Howard[1], MD; Tami Crawford[1], PhD; Lewis J Frey[1], PhD; Tracy Stecker[1], PhD; Brian E Bunnell[2], PhD

[1]Medical University of South Carolina, Charleston, SC, United States

[2]University of South Florida, Tampa, FL, United States

Corresponding Author:
Jihad S Obeid, MD
Medical University of South Carolina
135 Cannon St. Suite 405 MSC200
Charleston, SC, 29425
United States
Phone: 1 8437920272
Email: jobeid@musc.edu

Abstract

Background: Suicide is an important public health concern in the United States and around the world. There has been significant work examining machine learning approaches to identify and predict intentional self-harm and suicide using existing data sets. With recent advances in computing, deep learning applications in health care are gaining momentum.

Objective: This study aimed to leverage the information in clinical notes using deep neural networks (DNNs) to (1) improve the identification of patients treated for intentional self-harm and (2) predict future self-harm events.

Methods: We extracted clinical text notes from electronic health records (EHRs) of 835 patients with International Classification of Diseases (ICD) codes for intentional self-harm and 1670 matched controls who never had any intentional self-harm ICD codes. The data were divided into training and holdout test sets. We tested a number of algorithms on clinical notes associated with the intentional self-harm codes using the training set, including several traditional bag-of-words–based models and 2 DNN models: a convolutional neural network (CNN) and a long short-term memory model. We also evaluated the predictive performance of the DNNs on a subset of patients who had clinical notes 1 to 6 months before the first intentional self-harm event. Finally, we evaluated the impact of a pretrained model using Word2vec (W2V) on performance.

Results: The area under the receiver operating characteristic curve (AUC) for the CNN on the phenotyping task, that is, the detection of intentional self-harm in clinical notes concurrent with the events was 0.999, with an F1 score of 0.985. In the predictive task, the CNN achieved the highest performance with an AUC of 0.882 and an F1 score of 0.769. Although pretraining with W2V shortened the DNN training time, it did not improve performance.

Conclusions: The strong performance on the first task, namely, phenotyping based on clinical notes, suggests that such models could be used effectively for surveillance of intentional self-harm in clinical text in an EHR. The modest performance on the predictive task notwithstanding, the results using DNN models on clinical text alone are competitive with other reports in the literature using risk factors from structured EHR data.

(JMIR Med Inform 2020;8(7):e17784) doi:10.2196/17784

KEYWORDS

machine learning; deep learning; suicide; suicide, attempted; electronic health records; natural language processing

Introduction

Background and Significance

Suicide ranks among the leading causes of death in the United States. On average, over 100 individuals die of suicide each day, resulting in combined medical and work loss costs totaling approximately US $80 billion annually [1,2]. Numerous risk factors for suicide have been identified and thoroughly researched. For example, suicide is more common in males, American Indian and Alaska Natives, and non-Hispanics and individuals with mental illness (eg, depression, anxiety,

substance abuse), previous trauma, communication difficulties, decision-making impulsivity, and aggression [3,4]. Individuals who have previously engaged in intentional self-harm behaviors or suicide attempts are also at increased risk [5,6]. Despite extensive research on various risk factors, prospective suicide prediction remains difficult, as conventionally studied risk factors predict suicide attempts only 26% of the time [5].

Currently established guidelines for suicide risk assessment include clinical interviews and questionnaires administered by qualified health care providers [7,8]. However, research suggests that these approaches exhibit suboptimal performance in predicting future intentional self-harm behavior or suicide [9-11]. Less than a third of patients who engage in intentional self-harm and attempt suicide disclose thoughts about doing so [12]. As such, current methods for identification of at-risk patients can be difficult and time-consuming. A great deal of recent research has focused on addressing these limitations using advanced analytical tools such as natural language processing (NLP) and machine learning [13]. Studies using NLP approaches have largely used electronic health record (EHR)-based [14-16] and NLP- and linguistics-driven prediction models [12,17-19]. Studies using machine learning to predict suicidal and intentional self-harm behaviors from EHR data for patients admitted to hospitals or emergency departments have demonstrated variable accuracy (eg, 65%-95%) [20-24].

Clinical text classification using a deep convolutional network has been useful in the identification of specific phenotypes within the EHR for patients with a given set of clinical signs and symptoms [25,26]. There have been significant advances in recent years in deep learning approaches, such as convolutional neural networks (CNNs), for a variety of applications including text processing and classification, computer vision, and speech recognition [27]. In the area of text processing, there has been significant research in language models that are pretrained and then used to aid in automated text understanding of unlabeled data [28,29]. These resulting learned word vectors could, in turn, be used for clinical text classification tasks [25,26,30]. Pretraining models using these methods provide syntactic and semantic word similarities expressed in a multidimensional vector space with the potential for improving classifications based on neural networks and reducing computational cost [28]. The use of advanced analytical approaches such as deep learning can extend this work and provide distinct advantages in predicting future intentional self-harm, suicide attempts, and suicide.

Objectives

Deep learning approaches have been used to address topics related to suicide using publicly available data sets. For example, Shing et al [31] compared different machine learning methods including support vector machines (SVM) and a CNN-based model for the assessment of suicide risk based on web-based postings. Although they demonstrated the utility of deep learning, for this specific use case, the SVM model outperformed the CNN model. Conversely, Du et al [32] demonstrated the superiority of a deep learning model over traditional models, including an SVM, in identifying suicide-related tweets in social media data. Despite these examples, there have been no reports in the literature on the utility of deep learning approaches for the identification of suicide-related clinical records (eg, for surveillance purposes) or for the prediction of suicidal behavior using clinical text from an EHR. Improving the recall and precision of phenotyping and predictive algorithms, particularly through deep learning analytic techniques, could lead to better follow-up and care by clinicians for patients who are at risk for intentional self-harm, suicide attempts, suicide, or any combination thereof. In this study, we explored a deep learning approach for (1) the automated detection of intentional self-harm events in clinical text concurrent with International Classification of Diseases (ICD) codes for intentional self-harm, that is, phenotyping and (2) the prediction of future suicide attempts or intentional self-harm based on ICD-labeled encounters within the EHR.

Methods

Software Used

We used R version 3.6.1 (R Foundation for Statistical Computing) [33] for processing the data and clinical text and constructing the machine learning pipelines and Keras and TensorFlow v1.13 (Google's open-source deep neural network framework) for the deep learning models.

Patient Population

This study was approved by the institutional review board (IRB) for human research at the Medical University of South Carolina (MUSC) under protocol number Pro00087416. Clinical notes were extracted from the Epic (Epic Systems Corporation) EHR system [34] using the MUSC research data warehouse (RDW), which serves as an EHR data repository for research projects. Researchers may request data from the RDW with appropriate IRB approval and data governance oversight [35]. We extracted clinical text notes for adult patients aged 20 to 90 years with ICD codes for suicide attempts or intentional self-harm as defined in the National Health Statistics Report (NHSR) from the Centers for Disease Control and Prevention (CDC) in the United States [36]. The NHSR specifically included codes for self-harm events that were intentional (eg, T42.4X2; poisoning by benzodiazepines, intentional self-harm) and did not include codes for self-harm events that were unintentional (eg, T42.4X1; poisoning by benzodiazepines, accidental). For each patient in the study group, we selected the first intentional self-harm recorded in the chart during the study period (ie, 2012-2019). We filtered the notes within a 24-hour period of the intentional self-harm time stamp. We also extracted clinical text notes for control cases who never had any intentional self-harm ICD codes within our EHR spanning the years 2012 to 2019. The controls were selected randomly from the RDW after matching by age, gender, race, and ethnicity. During the processing of the clinical notes, we matched the controls to the study cases based on the proportion of note types in their records (eg, percent of progress notes) and word length of notes. The matching was performed using the nearest neighbor method in the MatchIt package in R [37]. The resulting patient population included 835 intentional self-harm cases and 1670 controls.

Clinical Notes

Notes Concurrent With Intentional Self-Harm

In the first part of this study, we sought to automate the detection of concurrent intentional self-harm ICD code assignment based on clinical text. The notes included a variety of different note types; however, the majority consisted of progress notes, plan of care notes, emergency department (ED) provider notes, history and physical (H&P) notes, and consult notes. A full list of note types and their relative frequencies is provided in a table in Multimedia Appendix 1. Individual notes longer than 800 words (less than one-third of all notes) were truncated at 800. We chose this cutoff to include as many notes per patient as possible. Notes belonging to the same patient were then concatenated into a single string arranged temporally, yielding 1 record per patient. Concatenated strings longer than 8000 words (44/2505, 1.76% of patients) were truncated at 8000. This allowed us to maintain the generated token vectors within a reasonable range for computational performance. The patients were divided into a training and cross-validation set (2012-2017) with 661 intentional self-harm cases and 1502 controls and a holdout test set (2018-2019) with 174 intentional self-harm cases and 168 controls.

Prediction From Previous Clinical Notes

In this part of the study, we sought to predict the future occurrence of intentional self-harm events based on previous clinical notes within the EHR. Clinical text was collected from a predictive window for a period between 180 days to 30 days before the index event (ie, the first reported intentional self-harm event on record) for each patient. Patients who did not have clinical notes during that time window were excluded. Clinical notes were used from the first date within that time window up to 90 days following the first date or up to 30 days before the intentional self-harm event (whichever is first). That is, the largest possible predictive window included clinical notes from a time interval of up to 90 days. The same time window was used for the control group; however, the latest visit on record within the study period was used as the index visit instead of an intentional self-harm event. To reduce noise and excessive amounts of notes in this part of the study, we limited notes to the following note types: progress notes, ED provider notes, H&P notes, consult notes, and discharge summaries. Individual notes were truncated to 1500 words and concatenated texts to 10,000-word cutoffs to capture a wider set of clinical texts. For the prediction part of the study, the patients were divided into a training and cross-validation set (2012-2017) with 480 intentional self-harm cases and 645 controls and a holdout test set (2018-2019) with 106 intentional self-harm cases and 106 controls.

Labeling the Test Set

A sample of 200 records from the test set (2018-2019) was manually reviewed to provide gold standard labels for a comparison with ICD code labels (based on the NHSR from the CDC). Each record reflected clinical notes in the EHR from concurrent visits of patients. We selected a random 100 from the study group (with intentional self-harm ICDs) and 100 controls. The concatenated strings from concurrent notes for this sample were imported into REDCap (Research Electronic Data Capture) [38] and made available for review and labeling by the reviewers on our research team, which included 3 clinical psychologists, a psychiatry resident, a medical student, and a pediatrician. The reviewers were instructed to label the notes as intentional self-harm if there was a suicide attempt or intentional self-harm noted in any of the clinical notes associated with the concurrent visit. Suicidal ideation alone was not considered intentional self-harm. A subsample of 100 notes was labeled independently by 2 labelers to estimate the interrater reliability.

Text Processing

We tested several machine learning algorithms using the training data, including both deep learning–based classifiers using word embeddings (WEs) and the traditional bag-of-words (BOW)–based models. We performed the necessary preprocessing of the text for both types. We used the quanteda R package [39] and regular expression functions within R for the text-processing pipeline. For the traditional BOW models, text processing included lower casing; removal of punctuation, stop words, and numbers; word stemming; and tokenization. For the WE models, text processing included lower casing, sentence segmentation, removal of punctuation, replacement of large numbers and dates with tokens using regular expressions, and tokenization.

Word Frequencies

Before running the machine learning algorithms, we examined differences in word frequencies across clinical notes concurrent with intentional self-harm events and notes preceding intentional self-harm events by over 30 days as compared with clinical notes from the control population. We performed a chi-square analysis to assess keywords that are overrepresented across the corpora of text [40].

Bag-of-Words−Based Classifiers

For the BOW models, word frequencies were used as features and were normalized using term frequency–inverse document frequency [41]. The traditional text classification models included naïve Bayes [42]; decision tree classifier [43] with a maximum depth of 20; random forest (RF) [44] with 201 trees and the number of variables randomly sampled as candidates at each split (mtry=150); SVM [45] type 1 with a radial basis kernel [46]; and a simple multilayer perceptron (MLP) artificial neural network with a 64-node input layer, a 64-node hidden layer, and a single output node. We used the rectified linear unit (ReLU) activation function in both the input and hidden layers and sigmoid activation for the binary output node. The MLP was trained using a learning rate of 1×10^{-4}, a batch size of 32, and a 20% validation split over 30 epochs.

Word Embeddings

We used Keras [47] and TensorFlow version 1.13 [48] for constructing and training the deep learning models. In preparation for WE, the text strings were converted to token sequences. To construct the features for the deep learning models, the sequences were prepadded with zeros to match the length of the longest string in the training set. We used

Word2vec (W2V) to generate a pretrained model [28]. The W2V weights were derived by pretraining a W2V skip-gram model on a sample of over 800,000 clinical notes from our EHR data set using 200 dimensions per word, a skip window size of 5 words in each direction, and negative sampling of 5. To explore and visualize the outcome of the pretrained W2V model, we used the t-distributed stochastic neighbor embedding (t-SNE) to map the multidimensional word vectors into a 2D space [49]. The performance of each deep learning classifier was assessed with either randomly initialized embeddings or W2V-initialized embeddings.

Deep Learning Models

We examined 2 different deep neural network (DNN) architectures: a CNN architecture similar to a previously published model [26] and a long short-term memory (LSTM) model [50]. Both architectures were tested using either randomly initialized WE weights in Keras or WE initialized with the weights from the pretrained W2V.

Both models had WE with 200 dimensions per word. The input layer had a dimension size slightly exceeding the maximum length of the input sequences of tokens, which were 8352 tokens for the concurrent notes and 11,000 tokens for the predictive notes. The CNN architecture consisted of an input layer; a WE layer included with a drop rate of 0.2; a convolutional layer with multiple filter sizes (3, 4, and 5) in parallel, with 200 nodes in each, ReLU activation, a stride of one, and global max-pooling; a merge tensor then a fully connected 200-node hidden layer with ReLU activation and a drop rate of 0.2; and an output layer with a single binary node with a sigmoid activation function. The LSTM architecture consisted of an input layer; a WE layer with a drop rate of 0.1; an LSTM layer with 64 nodes; both global average pooling and global max-pooling layers with a merge tensor of the 2; a fully connected 100-node hidden layer with ReLU activation and a drop rate of 0.1; and a single sigmoid binary output node.

The DNN models were trained using an adaptive moment estimation gradient descent algorithm [51] with a diminishing learning rate starting at 4×10^{-4}, batch size of 32, validation split at 15%, and early stopping based on the loss function for the validation data with patience of 5.

Training and Evaluation

Detection of Concurrent Intentional Self-Harm

For the automated detection of concurrent intentional self-harm ICD code assignment based on clinical text, we used the training and cross-validation data set (with index visits from 2012-2017) to identify the best performing models and hyperparameters. We then used the top 2 performing models (the DNNs) for training on the full training set and testing on the holdout test set (with index visits from 2018 to 2019), which included the 200 manually reviewed cases. The models were trained using intentional self-harm ICD codes as positive labels. However, we tested the output using both intentional self-harm ICD codes as positive labels and manually reviewed (gold standard) labels.

Prediction of Future Intentional Self-Harm Events

The 2 best performing models, namely, the DNNs, were used to predict future intentional self-harm events based on previous clinical notes. In the holdout test set, we used a balanced set with an equal number of intentional self-harm cases and controls with 106 cases in each. The DNN models were trained on notes preceding the first intentional self-harm visits during the 2012 to 2017 time frame and then tested on notes preceding the first intentional self-harm visits during the 2018 to 2019 time frame. Unlike the previous task, which had near-ceiling performance results with little variation, the performance of the DNNs on the predictive task varied between different runs of the same model even when using the same training and testing sets. This is due to the random initialization of weights in TensorFlow and random shuffling between epochs during training. To evaluate the performance of the different DNN architectures more precisely, we ran each model 50 times and examined the averages of the different metrics and used the Student *t* test (two-tailed) to determine statistical differences in performance.

Metrics

The performance metrics for all experiments, including area under the receiver operating characteristic (ROC) curve (AUC), were calculated in R using the caret [52] and pROC [53] packages. We also calculated the accuracy, precision, recall, and F1 score for all the models.

Results

International Classification of Diseases Code Analysis

The interrater reliability during the manual review exhibited a Cohen kappa of 0.96. Using the labels from the manual review as the gold standard, the accuracy of the intentional self-harm ICD codes attributed to concurrent visits was 0.92, with a precision of 0.84 and recall of 1.0. Thus, 16 cases out of 100 that were assigned an intentional self-harm ICD code did not exhibit intentional self-harm as part of the presenting history, per the manual review. However, all but 2 of the 16 *false-positives* by ICD had past intentional self-harm mentioned in their clinical notes. For those 2, 1 was suspected intentional self-harm, and the other had a previous admission for suicidal ideation with possible intentional self-harm.

Word Frequency Results

The result from this analysis overrepresented keywords in clinical notes concurrent with intentional self-harm events and clinical notes before the intentional self-harm events (Table 1). For example, the words *suicide* and *attempt* top the list in concurrent notes; however, they do not rank in the top 10 words in preceding notes. Instead, the words *disorder* and *si* (the shorthand for suicidal ideation) top the list in notes preceding intentional self-harm.

Table 1. The top 10 words in each group were compared with controls, along with the chi-square statistic for each.

Concurrent with ISH[a,b]		Before ISH[c]	
Keyword	Chi-square (*df*=1)	Keyword	Chi-square (*df*=1)
suicide	1.3E+5	disorder	1.2E+4
attempt	8.2E+4	si[d]	8.5E+3
overdose	6.7E+4	suicidal	6.0E+3
si	6.5E+4	mood	5.8E+3
disorder	5.2E+4	use	4.7E+3
suicidal	5.2E+4	alcohol	4.6E+3
psychiatry	4.0E+4	qhs[e]	4.5E+3
iop[f]	3.6E+4	safety	4.2E+3
interview	3.5E+4	interview	3.9E+3
mood	2.9E+4	cocaine	3.9E+3

[a]Keywords from clinical notes from visits concurrent with ISH events.

[b]ISH: intentional self-harm.

[c]Keywords from clinical notes from visits before the first ISH events.

[d]si: suicidal ideation.

[e]iop: Institute of Psychiatry.

[f]qhs: every bedtime (from Latin quaque hora somni).

Word2vec Pretraining Results

The W2V model successfully clustered words that seemed to have similar semantic contexts. Figure 1 shows the visualization of a sample of relevant words reduced into 2 dimensions using the t-SNE algorithm. Table 2 shows the top 10 words semantically similar to *attempt* and the top 10 words similar to *ideation* along with their cosine similarities. For example, the cosine similarity between *attempt* and *suicide* WE vectors was 0.730 and between *ideation* and *suicidal* was 0.872. The list also shows several misspelled words in a similar dimension space as their correctly spelled counterparts.

Figure 1. A visualization of a sample of relevant words derived from the Word2vec model reduced into two dimensions using t-distributed stochastic neighbor embedding. V1=variable 1; V2=variable 2.

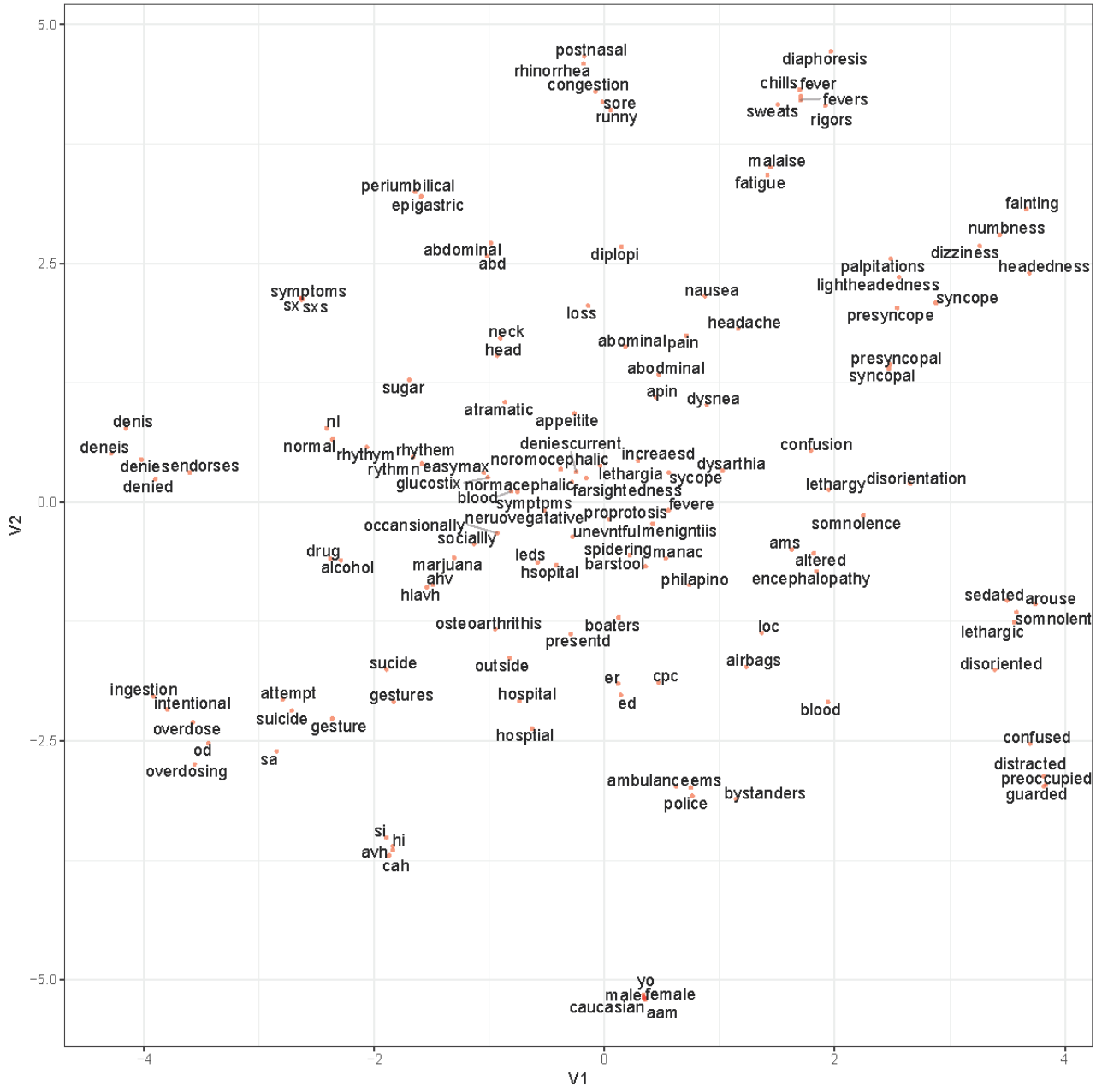


Table 2. Words semantically similar to the words attempt and ideation and their cosine similarity in the 200-dimension vector space as identified by the Word2vec analysis.

Term	Cos sim[a]
attempt	
attempt	1.000
suicide	0.730
overdose	0.696
osteoarthrithis	0.679
gesture	0.643
sucicide	0.625
benzodiaspines	0.619
intentional	0.617
ideation	Cos sim[a]
ideation	1.000
suicidal	0.872
homicidal	0.837
ideations	0.736
intent	0.681
ideaiton	0.651
si[b]	0.648
sucidial	0.619

[a]Cos sim: cosine similarity.

[b]si: suicidal ideation.

Detection of Concurrent Intentional Self-Harm

Training and Cross-Validation

Table 3 shows the results of the automated detection of concurrent intentional self-harm ICD code assignment based on the training and cross-validation data set with intentional self-harm visits during the period of 2012 to 2017. The DNNs outperformed the BOW classifiers. The CNN models had the highest AUC and F1 score. The best performance overall was for the CNN with W2V WE (CNNw) with an AUC of 0.988 and an F1 score of 0.928. The CNN with randomly initialized WE (CNNr) was a close second, with significantly overlapping 95% CIs. The LSTMs with randomly initialized WE (LSTMr) and the LSTM with W2V WE (LSTMw) AUCs were 0.982 and 0.975, respectively, with F1 scores above 0.887.

Among the BOW models, RF had the best AUC (0.961), and MLP had the best F1 score (0.862). On the basis of these results, we used 2 deep learning models for the rest of this study.

Table 3. The metrics for training and cross-validation on the 2012 to 2017 data set.

Model	AUC[a] (95% CI[b])	Accuracy (95% CI)	Precision	Recall	F1 score
NB[c]	0.908 (0.882-0.934)	0.870 (0.839-0.898)	0.734	0.865	0.794
DT[d]	0.870 (0.839-0.901)	0.865 (0.833-0.893)	0.715	0.885	0.791
RF[e]	0.961 (0.944-0.978)	0.896 (0.867-0.921)	0.794	0.865	0.828
SVM[f]	0.947 (0.925-0.969)	0.900 (0.872-0.924)	0.859	0.782	0.819
MLP[g]	0.957 (0.938-0.976)	0.917 (0.890-0.939)	0.828	0.897	0.862
CNNr[h]	0.984 (0.972-0.995)	0.946 (0.924-0.964)	0.938	0.872	0.904
CNNw[i]	0.988 (0.977-0.999)	0.959 (0.939-0.974)	0.947	0.910	0.928
LSTMr[j]	0.982 (0.972-0.992)	0.943 (0.920-0.961)	0.919	0.878	0.898
LSTMw[k]	0.975 (0.960-0.990)	0.937 (0.913-0.956)	0.918	0.859	0.887

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

[b]CI: 95% confidence intervals for the AUC.

[c]NB: naïve Bayes.

[d]DT: decision tree.

[e]RF: random forest.

[f]SVM: support vector machine.

[g]MLP: multilayer perceptron.

[h]CNNr: convolutional neural network with randomly initialized word embeddings.

[i]CNNw: convolutional neural network with Word2vec word embeddings.

[j]LSTMr: long short-term memory with randomly initialized word embeddings.

[k]LSTMw: long short-term memory with Word2vec word embeddings.

Testing of Concurrent Intentional Self-Harm Labels

Training the models on the full 2012 to 2017 data set then testing on the holdout (2018-2019) test set yielded even better performance than in the above cross-validation for detecting concurrent intentional self-harm ICD labels (Table 4). The best performing model was the CNNr with an AUC of 0.999 and an F1 score of 0.985. A plot of the training history for this task shows that the model converges smoothly to a minimum loss value on both training and validation (Multimedia Appendix 2). There was no advantage to adding the pretrained W2V WE, that is, the CNNw when testing on the holdout set. The CNNs slightly outperformed the LSTMs, but the results in all models were close to ceiling, making it difficult to point out the significance of these differences. As expected, as the models were trained on ICD labels, they performed better in predicting concurrent ICD labels than they did with predicting the gold standard labels (Figure 2). Of note, is that the recall remained very high when testing on the gold standard labels compared with the ICD labels, whereas the precision suffered slightly reflecting the precision achieved during the intentional self-harm ICD code analysis.

Table 4. The metrics for training on the 2012 to 2017 data set and testing on the 2018 to 2019 holdout test set using both International Classification of Diseases labels and gold standard labels.

Model	AUC[a] (95% CI[b])	Accuracy (95% CI)	Precision	Recall	F1 score
ICD[c] labels					
CNNr[d]	0.999 (0.998-1.000)	0.985 (0.957-0.997)	0.980	0.990	0.985
CNNw[e]	0.998 (0.996-1.000)	0.970 (0.936-0.989)	0.980	0.960	0.970
LSTMr[f]	0.997 (0.991-1.000)	0.980 (0.950-0.995)	0.990[d]	0.970	0.980
LSTMw[g]	0.997 (0.994-1.000)	0.960 (0.923-0.983)	0.989	0.930	0.959
Gold standard labels					
CNNr[c]	0.981 (0.966-0.997)	0.915 (0.867-0.950)	0.832	1.000	0.908
CNNw[e]	0.981 (0.965-0.997)	0.920 (0.873-0.954)	0.847	0.988	0.912
LSTMr[f]	0.968 (0.946-0.989)	0.910 (0.861-0.946)	0.837	0.976	0.901
LSTMw[g]	0.967 (0.945-0.989)	0.920 (0.873-0.954)	0.862	0.964	0.910

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

[b]CI: 95% confidence intervals for the AUC.

[c]ICD: International Classification of Diseases.

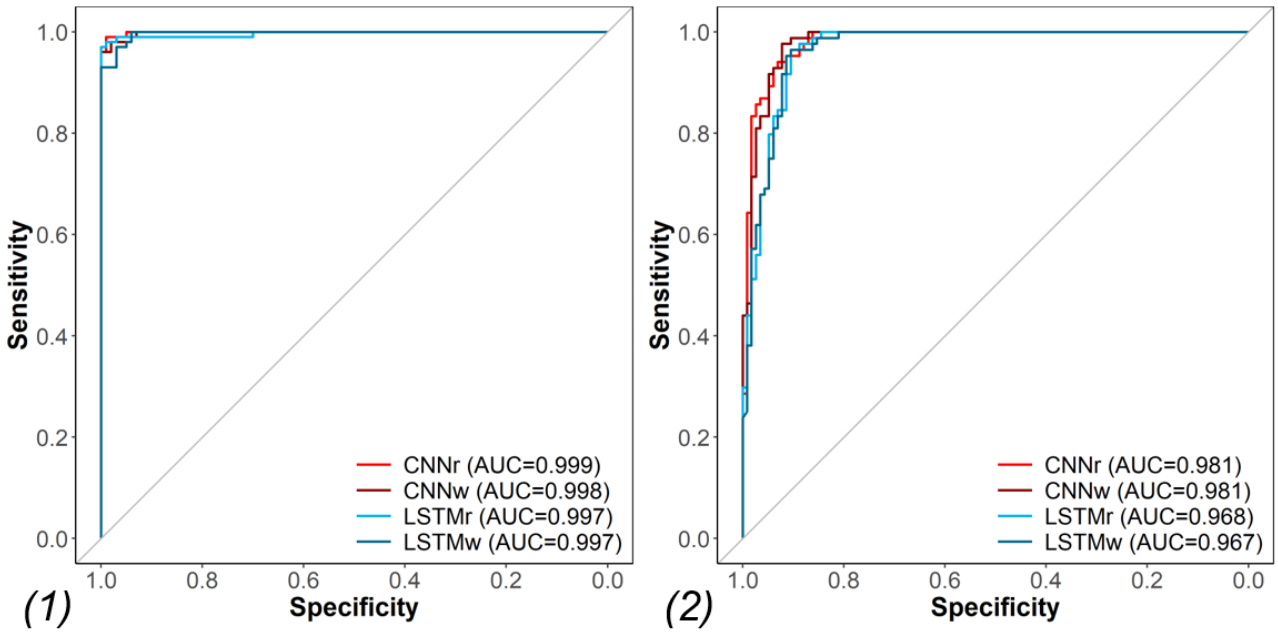
[d]CNNr: convolutional neural network with randomly initialized word embeddings.

[e]CNNw: convolutional neural network with Word2vec word embeddings.

[f]LSTMr: long short-term memory with randomly initialized word embeddings.

[g]LSTMw: long short-term memory with Word2vec word embeddings.

Figure 2. The area under the receiver operating characteristic curve for training on the 2012 to 2017 data set and testing on the holdout test set (2018-2019) using (1) International Classification of Diseases labels and (2) gold standard labels. AUC: area under the receiver operating characteristic curve; ICD: International Classification of Diseases; CNNr: convolutional neural network with randomly initialized word embeddings; CNNw: convolutional neural network with Word2vec word embeddings; LSTMr: long short-term memory with randomly initialized word embedding; LSTMw: long short-term memory with Word2vec word embedding.



Prediction of Future Intentional Self-Harm Events

The results for the prediction of future intentional self-harm events based on previous clinical notes are shown in Table 5. These values are the means of the different metrics after 50 training and testing cycles for each model. Figure 3 shows the differences in performance between the different models. The CNNr model had the best performance, with a mean AUC of 0.882 and a standard deviation of 0.006 (P<.001) compared with CNNw, which in turn outperformed the LSTM models

(*P*<.001). There was no significant difference between LSTMr and LSTMw. The variance in performance was notably wider in the LSTM models than in the CNN models. Multimedia Appendix 3 shows the ROC curves for each of the models highlighting the mean AUC. Although pretraining with W2V did not add value in terms of performance, it did reduce the number of epochs needed during training by an average of 32% for the CNN and 12% for the LSTM.

Table 5. The metrics for models trained on notes preceding the first intentional self-harm visits in patients presenting during the 2012 to 2017 time frame and tested on notes preceding the first intentional self-harm visits in patients presenting during the 2018 to 2019 time frame.

Model	AUC[a] (95% CI[b])	Accuracy (95% CI)	Precision	Recall	F1 score
CNNr[c]	0.882 (0.871-0.891)	0.792 (0.774-0.807)	0.863	0.694	0.769
CNNw[d]	0.869 (0.858-0.879)	0.782 (0.766-0.792)	0.860	0.673	0.755
LSTMr[e]	0.850 (0.827-0.877)	0.758 (0.729-0.788)	0.830	0.656	0.729
LSTMw[f]	0.846 (0.819-0.871)	0.750 (0.717-0.778)	0.822	0.644	0.720

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

[b]CI: 95% confidence intervals for the AUC.

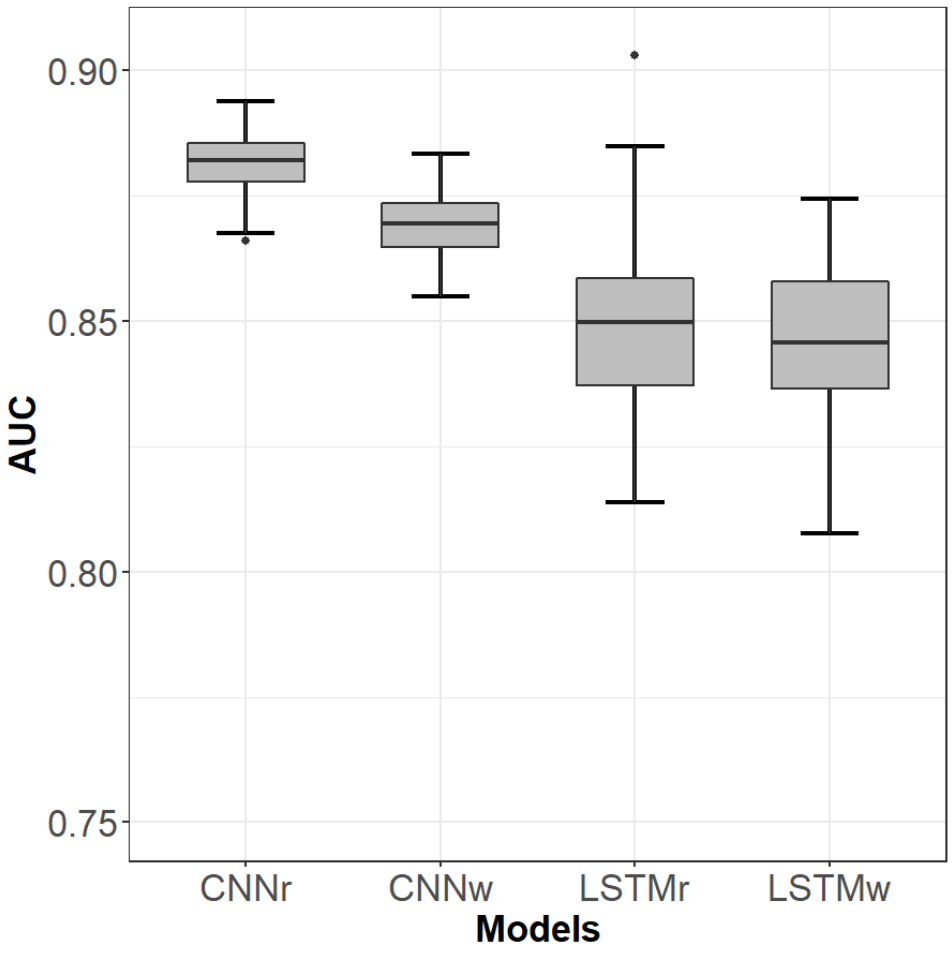
[b]CNNr: convolutional neural network with randomly initialized word embeddings.

[d]CNNw: convolutional neural network with Word2vec word embeddings.

[e]LSTMr: long short-term memory with randomly initialized word embeddings.

[f]LSTMw: long short-term memory with Word2vec word embeddings.

Figure 3. The mean area under the receiver operating characteristic curve and 95% CI for models trained on notes preceding the first intentional self-harm visits in patients presenting during the 2012 to 2017 time frame and tested on notes preceding the first intentional self-harm visits in patients presenting during the 2018 to 2019 time frame. The differences in performance were all significant (*P*<.001) except for the difference between the LSTMr and LSTMw. AUC: area under the receiver operating characteristic curve; CNNr: convolutional neural network with randomly initialized word embeddings; CNNw: convolutional neural network with Word2vec word embeddings; LSTMr: long short-term memory with randomly initialized word embedding; LSTMw: long short-term memory with Word2vec word embedding.



Discussion

Semantic Differences

The word frequency analyses identified keywords that were overrepresented in clinical notes associated with intentional self-harm visits. As noted in Table 1, words such as *attempt* and *overdose* were highly overrepresented in clinical notes concurrent with intentional self-harm events compared with controls. Conversely, suicidal ideation (as represented by the shorthand word *si*) was frequently present in preintentional self-harm notes. This is consistent with the literature on ideation, which is a prominent risk factor for suicide attempts and completions [54].

The W2V pretraining on our full data set of clinical notes successfully clustered relevant words together. It also demonstrated word similarity for some of the significant words identified above. For example, the words *attempt*, *suicide*, and *overdose* were closely linked with high cosine similarity. This model was also useful in clustering misspelled words with their correctly spelled counterparts, which may help reduce noise due to misspelling in the clinical notes.

Detection of Intentional Self-Harm Events

The deep learning models outperformed BOW models in identifying intentional self-harm in training and testing using the 2012 to 2017 data set. Given this outcome, we trained the deep learning models on the full 2012 to 2017 data set and then used the 2018 to 2019 data set as a holdout test set. This temporal division of the data is intended to replicate a real-world scenario where models could be trained on historical data to identify intentional self-harm in new records. The results show that we can accurately detect intentional self-harm events in concurrent clinical notes with intentional self-harm ICD codes. More specifically, we showed that a model trained on aggregated clinical text associated with a given intentional self-harm visit may be used to identify concurrent intentional self-harm events even if ICD codes were not yet provided or assigned. In other words, clinical text alone is useful in accurately identifying the intentional self-harm phenotype.

Although there is limited literature on the performance of NLP and machine learning approaches for the phenotyping of intentional self-harm, our DNN classifiers with precisions up to 99% for concurrent notes with intentional self-harm ICD codes and up to 86% for gold standard intentional self-harm events compare favorably with previous reports, especially when considering that the models were trained on ICD codes as labels. Using a hybrid machine learning and rule-based NLP approach, Fernandes et al [19] achieved a precision of 82.8% for identifying suicide attempts. Another study comparing the accuracy of ICD codes and NLP-extracted concepts for suicidality achieved a precision of 60% using NLP alone and 97% using both ICD-9 codes and NLP; however, this study did not differentiate between suicidal ideation and intentional self-harm [16].

Although the CNN-based models seemed to slightly outperform the LSTM-based models on the phenotyping task, it is difficult to show a significant advantage to using either model or the advantage of pretraining with W2V due to the near-ceiling performance of all the DNNs on this task and the relatively small data set.

Nonetheless, a DNN model trained using this method may be useful for surveillance purposes and could well supplement surveillance using ICD codes. Training such a model using intentional self-harm ICD codes as positive labels is dependent on reliable assignment of ICD codes. Fortunately, ICD codes for intentional self-harm at our institution were accurate, as shown by the manual review of charts, notwithstanding the limitation of a relatively high false-positive rate. Finally, accurate phenotyping of the intentional self-harm events paves the way for future directions in identifying other phenotypes, for example, those with suicidal ideation alone versus intentional self-harm or not intentional self-harm, which may or may not have accurate ICD codes. Such precise or deep phenotyping is an important step toward predicting the risk of mortality, given the availability of mortality data.

Prediction of Future Intentional Self-Harm Based on Clinical Text

The results also show that aggregated clinical notes from visits between 1 and 6 months before the index visit predicted future intentional self-harm events with an AUC of 0.882 for the *best performing* CNN model. These results compare favorably with the literature on predictive models for suicide attempts. Using a complex combination of structured EHR data (including demographics, diagnostic codes, and census-based socioeconomic status) and medication data extracted via NLP, Walsh et al [20] achieved a maximum AUC of 0.84. Moreover, this AUC was based only on 7-day-old data. The AUC dropped gradually to 0.81 as the predictive window widened to 6 months before the index visit.

When comparing the performance between the 2 DNN architectures, we noted a consistent and statistically significant performance advantage of the 2 CNN models over the LSTM-based ones (Figure 3). Moreover, the LSTM had a relatively high variance and inconsistent performance over the 50 training runs, as can be noted from the CIs. We also noted a higher computational cost for the LSTM over the CNN (almost twice the time needed for training per epoch). In addition to the higher computational cost, recurrent neural networks show a minor advantage in generic text classification tasks [55,56]. At least with a small data set like ours, the CNNs were found to converge more smoothly and provide better performance.

While the W2V pretraining clustered similar words, initializing the WE layer with W2V weights did not add any value to either of the predictive models. Although CNNr (AUC=0.882) performed only slightly better than CNNw (AUC=0.869), the difference was statistically significant. However, there was no difference between the LSTMr and LSTMw. These results were unexpected given the advantages of pretrained WE in picking up misspellings and word similarities and highlight the need to examine newer, more complex language models such as Google's (Alphabet Inc) Bidirectional Encoder Representations from Transformers [29].

Regardless of the model architecture, these results are promising. Such predictive models may be useful in stratifying hospitalized patients into risk categories, which may aid in discharge planning. Using technology (telephone, emails, or text messages) for follow-up in the postdischarge period has been shown to reduce risk of future suicide attempts [57]. Furthermore, patients could be prophylactically assigned a social worker; be directed to collaborative primary care clinics with access to mental health services; or receive mental health referrals, telehealth appointments, or home health visits [58]. Adequate refinement of a predictive model may even allow for stratification of patients to a level of care necessary post discharge, beyond simple binary risk categorization.

Limitations

To identify patients with intentional self-harm during a given visit, we trained the models on ICD codes. Therefore, they can only perform as well as the ICD code designation. As mentioned earlier, during the manual labeling process, several patients had a past medical history of intentional self-harm rather than suicide attempt or self-harm as part of the presenting chief complaint or diagnosis. A possible solution would be to train models to introduce multiple labels that include current and past intentional self-harm through manual review. However, this would require a manual review of several hundreds of charts, which was beyond the scope of this initial pilot work.

Moreover, although we can clearly identify intentional self-harm, this still does not specify *intent to die*. This highlights the need for data on fatalities due to suicide. There are multiple forms of self-injury (eg, firearms, sharp objects, jumping from a high place) with ICD codes that are not accompanied by the classification of intent to harm oneself. Therefore, in these instances of unknown intent, self-injury may reflect a multitude of motives: communicating distress, suicidal gestures with low lethality, nonsuicidal self-injury (NSSI), or fatality [59]. Existing literature predicting NSSI behaviors yields 3 notable risk factor categories: history of NSSI, cluster B personality, and hopelessness [60]. Identifying NSSI can be of a significant prognostic value and has not been distinguished from intent to die in this study.

Another limitation of this study is that our model currently only addresses features within clinical texts. Other clinical information could be added to the model, such as associated demographics, comorbidities, and risk factors (eg, codes for depression or substance use). Moreover, with respect to suicide prediction, EHR data alone may not provide a full picture. Ideally, our data should be linked with the statewide cause of death data, which should yield an improved predictive power.

Although deep learning models are more powerful, they are less interpretable than some of the BOW models. For example, when using an RF model, the results of a variable importance analysis may yield insight into significant words. In fact, it may be beneficial to use both types of predictive models in mental health applications. This would leverage the power of deep learning models as well as the advantages of interpretable models. Future work should also include the exploration of attention-based deep learning models with some insight into explainability [61], which may address the utility of these models in real-world clinical decision support and adoption by clinicians.

Finally, the results presented here are based on data from a single EHR system at 1 academic medical center, making it difficult to draw generalizations about the high level of performance of our models in other environments. Future work should include collaboration with other institutions to ascertain the performance of these models in other environments.

Conclusions

Most of the models showed relatively good performance when detecting intentional self-harm events in concurrent clinical notes, that is, the phenotyping task. This is likely due to a strong signal within concurrent notes and is associated with a high fidelity of ICD code attribution for intentional self-harm, at least at our institution. When applied to the prediction of a future occurrence of intentional self-harm code assignment in a patient chart based on previous clinical notes, the AUC dropped to 0.882 with a modest recall and precision. Nevertheless, our results are competitive with the results from other models reported in the literature. Improving the precision of these algorithms could lead to better follow-up and preventative care by mental health professionals for patients who are at risk for future suicide attempts.

Acknowledgments

This project was supported, in part, by the National Center for Advancing Translational Sciences of the National Institutes of Health under grant number UL1 TR001450, the National Institute on Drug Abuse (K23 DA045766 to JD), and the National Institute of Mental Health (K23 MH118482 to BB). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

JD is co-owner of the Behavioral Activation Tech LLC, a company that develops technology-based treatments for depression.

Multimedia Appendix 1
The full list of note types and their relative frequencies in the data set.
[DOCX File , 19 KB - medinform_v8i7e17784_app1.docx]

Multimedia Appendix 2
A plot of the convolutional neural network model's training history for the phenotyping task. The learning curve shows that the model converges smoothly to a minimum loss value on both training and validation sets using an Adam optimizer.
[PNG File , 37 KB - medinform_v8i7e17784_app2.png]

Multimedia Appendix 3
Plots of the receiver operating characteristic curves for the 50 training and testing runs for all the models highlighting the mean area under the receiver operating characteristic curve for each model.
[PNG File , 235 KB - medinform_v8i7e17784_app3.png]

References

1. Kochanek KD, Murphy S, Xu J, Arias E. Mortality in the United States, 2016. NCHS Data Brief 2017 Dec(293):1-8 [FREE Full text] [Medline: 29319473]
2. Fatal Injury Data. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/injury/wisqars/fatal.html [accessed 2019-12-02]
3. Preventing Suicide. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/violenceprevention/suicide/fastfact.html [accessed 2019-12-02]
4. Gvion Y, Levi-Belz Y. Serious suicide attempts: systematic review of psychological risk factors. Front Psychiatry 2018;9:56 [FREE Full text] [doi: 10.3389/fpsyt.2018.00056] [Medline: 29563886]
5. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
6. Chan MK, Bhatti H, Meader N, Stockton S, Evans J, O'Connor RC, et al. Predicting suicide following self-harm: systematic review of risk factors and risk scales. Br J Psychiatry 2016 Oct;209(4):277-283. [doi: 10.1192/bjp.bp.115.170050] [Medline: 27340111]
7. Jacobs D, Baldessarini R, Conwell Y, Fawcett J, Horton L, Meltzer H, et al. Assessment and treatment of patients with suicidal behaviors. Practice Guideline for the Assessment and Treatment of Patients with Suicidal Behaviors. Washington, D.C: American Psychiatric Association Steering Committee on Practice Guidelines; 2010. URL: https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/suicide.pdf [accessed 2020-06-10]
8. Surveillance Report 2016 – Self-Harm in Over 8s: Short-Term Management and Prevention of Recurrence (2004) NICE Guideline CG16 and Self-Harm in Over 8s: Long Term Management (2011) NICE Guideline CG133. London: National Institute for Health and Care Excellence (UK); 2016.
9. Larkin C, Di Blasi Z, Arensman E. Risk factors for repetition of self-harm: a systematic review of prospective hospital-based studies. PLoS One 2014;9(1):e84282 [FREE Full text] [doi: 10.1371/journal.pone.0084282] [Medline: 24465400]
10. Bolton JM, Gunnell D, Turecki G. Suicide risk assessment and intervention in people with mental illness. Br Med J 2015 Nov 9;351:h4978. [doi: 10.1136/bmj.h4978] [Medline: 26552947]
11. Runeson B, Odeberg J, Pettersson A, Edbom T, Adamsson IJ, Waern M. Instruments for the assessment of suicide risk: a systematic review evaluating the certainty of the evidence. PLoS One 2017;12(7):e0180292 [FREE Full text] [doi: 10.1371/journal.pone.0180292] [Medline: 28723978]
12. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, et al. Predicting the risk of suicide by analyzing the text of clinical notes. PLoS One 2014;9(1):e85733 [FREE Full text] [doi: 10.1371/journal.pone.0085733] [Medline: 24489669]
13. Burke TA, Ammerman BA, Jacobucci R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: a systematic review. J Affect Disord 2019 Feb 15;245:869-884. [doi: 10.1016/j.jad.2018.11.073] [Medline: 30699872]
14. Kessler RC, Stein MB, Petukhova MV, Bliese P, Bossarte RM, Bromet EJ, Army STARRS Collaborators. Predicting suicides after outpatient mental health visits in the army study to assess risk and resilience in servicemembers (army STARRS). Mol Psychiatry 2017 Apr;22(4):544-551 [FREE Full text] [doi: 10.1038/mp.2016.110] [Medline: 27431294]
15. Zhong Q, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, et al. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs clinical notes processed by natural language processing. BMC Med Inform Decis Mak 2018 May 29;18(1):30 [FREE Full text] [doi: 10.1186/s12911-018-0617-7] [Medline: 29843698]
16. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. AMIA Annu Symp Proc 2012;2012:1244-1253 [FREE Full text] [Medline: 23304402]
17. McCoy TH, Castro VM, Roberson AM, Snapper LA, Perlis RH. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry 2016 Oct 1;73(10):1064-1071. [doi: 10.1001/jamapsychiatry.2016.2172] [Medline: 27626235]

18. Downs J, Velupillai S, George G, Holden R, Kikoler M, Dean H, et al. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. AMIA Annu Symp Proc 2017;2017:641-649 [FREE Full text] [Medline: 29854129]
19. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Sci Rep 2018 May 9;8(1):7426 [FREE Full text] [doi: 10.1038/s41598-018-25773-2] [Medline: 29743531]
20. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science 2017 Apr 11;5(3):457-469. [doi: 10.1177/2167702617691560]
21. Delgado-Gomez D, Baca-Garcia E, Aguado D, Courtet P, Lopez-Castroman J. Computerized adaptive test vs decision trees: development of a support decision system to identify suicidal behavior. J Affect Disord 2016 Dec;206:204-209. [doi: 10.1016/j.jad.2016.07.032] [Medline: 27475891]
22. Metzger M, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. Int J Methods Psychiatr Res 2017 Jun;26(2):- [FREE Full text] [doi: 10.1002/mpr.1522] [Medline: 27634457]
23. Lopez-Castroman J, Perez-Rodriguez MD, Jaussent I, Alegria AA, Artes-Rodriguez A, Freed P, European Research Consortium for Suicide (EURECA). Distinguishing the relevant features of frequent suicide attempters. J Psychiatr Res 2011 May;45(5):619-625. [doi: 10.1016/j.jpsychires.2010.09.017] [Medline: 21055768]
24. Mann JJ, Ellis SP, Waternaux CM, Liu X, Oquendo MA, Malone KM, et al. Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making. J Clin Psychiatry 2008 Jan;69(1):23-31 [FREE Full text] [doi: 10.4088/jcp.v69n0104] [Medline: 18312034]
25. Turner CA, Jacobs AD, Marques CK, Oates JC, Kamen DL, Anderson PE, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC Med Inform Decis Mak 2017 Aug 22;17(1):126 [FREE Full text] [doi: 10.1186/s12911-017-0518-1] [Medline: 28830409]
26. Obeid JS, Weeda ER, Matuskowitz AJ, Gagnon K, Crawford T, Carr CM, et al. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. BMC Med Inform Decis Mak 2019 Aug 19;19(1):164 [FREE Full text] [doi: 10.1186/s12911-019-0894-9] [Medline: 31426779]
27. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
28. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. 2013 Jan 16. URL: https://arxiv.org/pdf/1301.3781.pdf [accessed 2018-11-20]
29. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. 2019 May 24. URL: https://www.aclweb.org/anthology/N19-1423.pdf [accessed 2020-06-10]
30. Banerjee I, Madhavan S, Goldman RE, Rubin DL. Intelligent word embeddings of free-text radiology reports. AMIA Annu Symp Proc 2017;2017:411-420 [FREE Full text] [Medline: 29854105]
31. Shing H, Nair S, Zirikly A, Friedenberg M, Daumé IH, Resnik P. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. 2018 Presented at: CLPsych'18; June 5, 2018; New Orleans, LA. [doi: 10.18653/v1/w18-0603]
32. Du J, Zhang Y, Luo J, Jia Y, Wei Q, Tao C, et al. Extracting psychiatric stressors for suicide from social media using deep learning. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):43 [FREE Full text] [doi: 10.1186/s12911-018-0632-8] [Medline: 30066665]
33. R: A Language and Environment for Statistical Computing. The R Project. 2019. URL: https://www.r-project.org/ [accessed 2019-12-27]
34. Epic. URL: https://www.epic.com/ [accessed 2019-06-05]
35. Obeid JS, Beskow LM, Rape M, Gouripeddi R, Black RA, Cimino JJ, et al. A survey of practices for the use of electronic health records to support research recruitment. J Clin Transl Sci 2017 Aug;1(4):246-252 [FREE Full text] [doi: 10.1017/cts.2017.301] [Medline: 29657859]
36. Hedegaard H, Schoenbaum M, Claassen C, Crosby A, Holland K, Proescholdbell S. Issues in developing a surveillance case definition for nonfatal suicide attempt and intentional self-harm using international classification of diseases, tenth revision, clinical modification (ICD-10-CM) coded data. Natl Health Stat Report 2018 Feb(108):1-19 [FREE Full text] [Medline: 29616901]
37. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 2017 Jan 04;15(3):199-236. [doi: 10.1093/pan/mpl013]
38. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
39. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, et al. quanteda: an R package for the quantitative analysis of textual data. J Open Source Softw 2018 Oct;3(30):774. [doi: 10.21105/joss.00774]

40. Culpeper J. Keyness: words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. Int J Corpus Linguist 2009;14(1):29-59. [doi: 10.1075/ijcl.14.1.03cul]
41. Manning C, Raghavan P, Schutze H. Introduction to Information Retrieval. Cambridge, USA: Cambridge University Press; 2008.
42. McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification. AAAI-98 Workshop 'Learning for Text Categorization'. 1998. URL: https://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf [accessed 2020-06-10]
43. Breiman L. Classification and Regression Trees. New York, USA: Chapman & Hall; 1984.
44. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
45. Weston J, Watkins C. Multi-class support vector machines. In: Support Vector Machines Applications. Switzerland: Springer International Publishing; 1998:-.
46. Joachims T. In: Nédellec C, Rouveirol C, editors. Text categorization with Support Vector Machines: Learning With Many Relevant Features. Berlin Heidelberg: Springer; 1998:E.
47. Chollet F. Keras. 2018. URL: https://keras.io/ [accessed 2018-11-20]
48. TensorFlow. 2018. URL: https://www.tensorflow.org/ [accessed 2018-11-20]
49. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579-2605 [FREE Full text]
50. Lee J, Dernoncourt F. Sequential short-text classification with recurrent and convolutional neural networks. arXiv. 2016 Mar 11. URL: https://www.aclweb.org/anthology/N16-1062.pdf [accessed 2020-06-10]
51. Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv. 2019. URL: https://arxiv.org/pdf/1412.6980.pdf [accessed 2020-06-10]
52. Kuhn M. The Caret Package. GitHub. URL: http://topepo.github.io/caret/index.html [accessed 2018-12-06]
53. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006 Jun;27(8):861-874. [doi: 10.1016/j.patrec.2005.10.010]
54. Klonsky ED, May AM, Saffer BY. Suicide, suicide attempts, and suicidal ideation. Annu Rev Clin Psychol 2016;12:307-330. [doi: 10.1146/annurev-clinpsy-021815-093204] [Medline: 26772209]
55. Vu NT, Adel H, Gupta P, Schütze H. Combining recurrent and convolutional neural networks for relation classification. Association for Computational Linguistics. 2016 May 24. URL: https://www.aclweb.org/anthology/N16-1065.pdf [accessed 2020-06-10]
56. Wen Y, Zhang W, Luo R, Wang J. Learning text representation using recurrent convolutional neural network with highway layers. University College London. 2016 Aug 2. URL: https://discovery.ucl.ac.uk/id/eprint/1526824/1/Wang_neuir2016.pdf [accessed 2020-06-10]
57. Falcone G, Nardella A, Lamis DA, Erbuto D, Girardi P, Pompili M. Taking care of suicidal patients with new technologies and reaching-out means in the post-discharge period. World J Psychiatry 2017 Sep 22;7(3):163-176 [FREE Full text] [doi: 10.5498/wjp.v7.i3.163] [Medline: 29043154]
58. Sall J, Brenner L, Millikan Bell AM, Colston MJ. Assessment and management of patients at risk for suicide: synopsis of the 2019 US department of veterans affairs and US department of defense clinical practice guidelines. Ann Intern Med 2019 Sep 3;171(5):343-353. [doi: 10.7326/M19-0687] [Medline: 31450237]
59. Heilbron N, Compton JS, Daniel SS, Goldston DB. The problematic label of suicide gesture: alternatives for clinical research and practice. Prof Psychol Res Pr 2010 Jun 1;41(3):221-227 [FREE Full text] [doi: 10.1037/a0018712] [Medline: 20640243]
60. Fox KR, Franklin JC, Ribeiro JD, Kleiman EM, Bentley KH, Nock MK. Meta-analysis of risk factors for nonsuicidal self-injury. Clin Psychol Rev 2015 Dec;42:156-167 [FREE Full text] [doi: 10.1016/j.cpr.2015.09.002] [Medline: 26416295]
61. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable prediction of medical codes from clinical text. arXiv. 2018 Apr 16. URL: https://arxiv.org/pdf/1802.05695.pdf [accessed 2020-06-10]

Abbreviations

AUC: area under the receiver operating characteristic curve
BOW: bag-of-words
CDC: Centers for Disease Control and Prevention
CNN: convolutional neural network
CNNr: CNN with randomly initialized word embedding
CNNw: CNN with Word2vec word embedding
DNN: deep neural network
ED: emergency department
EHR: electronic health record
H&P: history and physical
ICD: International Classification of Diseases
IRB: institutional review board
LSTM: long short-term memory

LSTMr: LSTM with randomly initialized word embedding
LSTMw: LSTM with Word2vec word embedding
MLP: multilayer perceptron
MUSC: Medical University of South Carolina
NHSR: National Health Statistics Report
NLP: natural language processing
NSSI: nonsuicidal self-injury
RDW: research data warehouse
ReLU: rectified linear unit
RF: random forest
ROC: receiver operating characteristic
SVM: support vector machines
t-SNE: t-distributed stochastic neighbor embedding
W2V: Word2vec
WE: word embedding

Edited by J Bian; submitted 13.01.20; peer-reviewed by F Li, S Shams; comments to author 05.03.20; revised version received 25.04.20; accepted 21.05.20; published 30.07.20.

Please cite as:
Obeid JS, Dahne J, Christensen S, Howard S, Crawford T, Frey LJ, Stecker T, Bunnell BE
Identifying and Predicting Intentional Self-Harm in Electronic Health Record Clinical Notes: Deep Learning Approach
JMIR Med Inform 2020;8(7):e17784
URL: https://medinform.jmir.org/2020/7/e17784
doi:10.2196/17784
PMID:32729840

Original Paper

An Ensemble Learning Strategy for Eligibility Criteria Text Classification for Clinical Trial Recruitment: Algorithm Development and Validation

Kun Zeng[1], PhD; Zhiwei Pan[1], MA; Yibin Xu[2], BA; Yingying Qu[3], PhD

[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[2]School of Computer Science, South China Normal University, Guangzhou, China
[3]School of Business, Guangdong University of Foreign Studies, Guangzhou, China

Corresponding Author:
Yingying Qu, PhD
School of Business
Guangdong University of Foreign Studies
178 Outer Ring East Road, Panyu District
Guangzhou, 510000
China
Phone: 86 15521018804
Email: jessie.qu@gdufs.edu.cn

Abstract

Background: Eligibility criteria are the main strategy for screening appropriate participants for clinical trials. Automatic analysis of clinical trial eligibility criteria by digital screening, leveraging natural language processing techniques, can improve recruitment efficiency and reduce the costs involved in promoting clinical research.

Objective: We aimed to create a natural language processing model to automatically classify clinical trial eligibility criteria.

Methods: We proposed a classifier for short text eligibility criteria based on ensemble learning, where a set of pretrained models was integrated. The pretrained models included state-of-the-art deep learning methods for training and classification, including Bidirectional Encoder Representations from Transformers (BERT), XLNet, and A Robustly Optimized BERT Pretraining Approach (RoBERTa). The classification results by the integrated models were combined as new features for training a Light Gradient Boosting Machine (LightGBM) model for eligibility criteria classification.

Results: Our proposed method obtained an accuracy of 0.846, a precision of 0.803, and a recall of 0.817 on a standard data set from a shared task of an international conference. The macro F1 value was 0.807, outperforming the state-of-the-art baseline methods on the shared task.

Conclusions: We designed a model for screening short text classification criteria for clinical trials based on multimodel ensemble learning. Through experiments, we concluded that performance was improved significantly with a model ensemble compared to a single model. The introduction of focal loss could reduce the impact of class imbalance to achieve better performance.

(JMIR Med Inform 2020;8(7):e17832) doi:10.2196/17832

KEYWORDS

Deep learning; Text classification; Ensemble learning; Eligibility criteria; Clinical trial

Introduction

Clinical trials are experiments or observations conducted on human volunteers, who are also referred to as subjects in clinical research. Eligibility criteria are the main indicators developed by those conducting the clinical trial to identify whether a subject should be enrolled in a clinical trial [1]. The criteria consist of inclusion and exclusion criteria, which are generally unstructured texts. Recruitment of subjects for clinical trials is generally conducted by manual comparison of their medical records with clinical trial eligibility criteria [2]. In 2009, Thadani et al [3] pointed out that manual comparison was time-consuming, labor-intensive, and inefficient compared with electronic screening. Therefore, clinical trials face many difficulties in recruitment, including difficulty in finding subjects and a long recruitment time [4]. Using natural language processing and machine learning methods to automatically

analyze clinical trial eligibility criteria texts and build an automated patient screening system is a promising research topic, with great practical application prospects and clinical value [5,6]. In 2016, Agarwal et al [7] proposed a model to predict the probability of users' future visits to a medical facility by constructing a matrix of semantic and location-based features from search logs of a search engine.

Text classification is an essential research topic in text information processing. It associates a given text with one or more categories based on characteristics of the text (content, attributes, or features), under a predefined classification taxonomy. Effective feature selection is crucial to the efficiency and accuracy of text classification tasks [8]. Using text classification technology to process medical texts, such as electronic medical records, not only improves the work efficiency of medical institutions [3], but also provides a basis for the further processing of medical text data. In addition, text classification technology has great significance for the research of knowledge graph construction [9], question answering system design [10], and automatic text summary [11].

However, unlike open domains, the complexity of medical texts makes it extremely difficult to classify them. First, the complexity of medical texts mainly comes from a large number of domain-specific terms. Different categories of texts correspond to medical terms of disease names, drug names, body part names, and other information, which presents difficulties in text segmentation and subsequent text feature extraction [12]. Second, the diversity of medical natural language texts also increases the complexity of medical text classification [13]. For example, a disease concept may have more than 10 mentions in a disease category. In addition, this type of medical text data is generally imbalanced, which presents difficulties in the classification of categories that contain a small amount of data [14].

With the rapid development of deep learning [15], many short text classification methods based on word vector models have emerged. Kaljahi et al [16] proposed the Any-gram kernel method to extract N-gram features of short texts, and used a bidirectional long-term and short-term memory network (BILSTM) to classify the texts. The method made improvements in topic- and sentence-level sentiment analysis tasks. Kim et al [17] used convolutional neural networks (CNN) to solve sentence classification problems. Lee et al [18] combined recurrent neural networks (RNN) and convolutional neural networks to classify short texts. Hsu et al [19] mixed convolutional neural networks with recurrent neural networks and proposed a structure-independent gate representation algorithm for sentence classification. Zhou et al [20] introduced a 2-dimensional maximum pooling operation to a bidirectional long-term and short-term memory network (BILSTM) to extract the features of texts in the temporal and spatial dimensions in a text classification task. In recent years, the Bidirectional Encoder Representations from Transformers (BERT) model [21] proposed by Google utilized a self-attention mechanism transformer [22], which improved feature extraction capability based on a long-term and short-term memory network (LSTM) and improved the bidirectional fusion function in stitching mode.

In order to solve the difficulties (eg, feature extraction) caused by a large number of domain specific diseases, medicines, body parts names, and other terminology, our paper proposed a character-level short text classification model. For word embedding, 4 character-level word embedding models were selected: BERT, A Robustly Optimized BERT Pretraining Approach (RoBERTa), XLNet, and Enhanced Representation through Knowledge Integration (ERNIE). We used a pretrained model based on Chinese corpus to accelerate the convergence of the model. In order to reduce the data imbalance problem, focal loss was introduced to the training process to train the model more stably. Finally, LightGBM was used to ensemble the 4 models to improve overall performance.

The main contributions of this paper are as follows: (1) a character-level ensemble learning model created by integrating BERT, RoBERTa, XLNet, and ERNIE was proposed for eligibility criteria text classification. (2) The focal loss as a loss function was leveraged to solve the problem of data imbalance among different categories. (3) The evaluation results showed that our ensemble learning model outperformed several baseline methods, demonstrating its effectiveness in the eligibility criteria text classification task.

Methods

Data Set

Our data set comes from the evaluation task of the China Health Information Processing Conference (CHIP) 2019. There are three evaluation tasks. The first task is the standardization of clinical terms [23]. The main goal of this task is to standardize the semantics of surgical entity mentions in Chinese electronic medical records. Given a surgical word, the corresponding standardized word is required. The second task is disease question transfer learning [24]. The main objective is to perform transfer learning between diseases based on Chinese disease question and answer data. Specifically, given question pairs in five different disease types, it is required to determine whether the semantics of two questions are the same or similar. The third evaluation task is the short text classification of clinical trial eligibility criteria.

The data set contains 38341 clinical trial eligibility criteria texts and has been manually annotated by human experts. Table 1 shows some specific examples of eligibility criteria texts and their annotated categories. For instance, the corresponding category of "血糖<2.7 mmol/L" (blood glucose<2.7 mmol/L) is "Laboratory Examinations."

The data set contains 44 various categories of clinical trial eligibility criteria in total, including "Disease," "Multiple," "Therapy or Surgery," etc. The data set is further divided into a training set, a validation set, and a test set. The training set contains 22962 pieces of eligibility criteria texts, while the validation and test sets contain 7682 and 7697 texts, respectively.

Table 1. Examples of eligibility criteria texts and corresponding annotated categories.

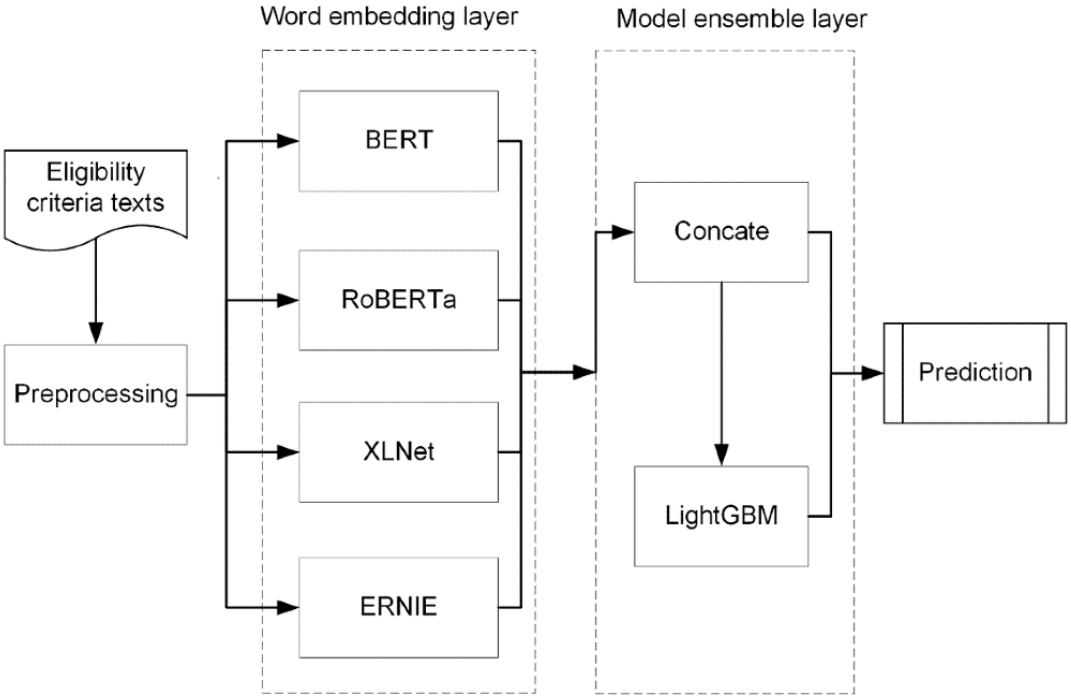
Eligibility criteria text	Annotated category
年龄>80岁 (Age>80)	Age
近期颅内或椎管内手术史 (Recent intracranial or spinal canal surgery)	Therapy or surgery
血糖<2.7 mmol/L (Blood glucose<2.7 mmol/L)	Laboratory examinations
2)性别不限,年龄18～70岁 (Unlimited gender, aged 18-70 years)	Multiple
合并造血系统或恶性肿瘤等严重原发性疾病 (A serious primary disease, such as one involving the hematopoietic system or a malignant tumor)	Disease
其他研究者认为不适合参加本研究的患者 (Patients that are unsuitable for this study that were considered by other investigators)	Researcher decision
预期生存超过12周 (Expected survival over 12 weeks)	Life expectancy
男、女不限 (Male or female)	Gender

Overall Framework

The overall framework of our proposed model is shown in Figure 1. As shown in the flowchart, the sample texts were preprocessed and converted from characters to numeric vectors for training. After that, we used BERT, RoBERTa, XLNet, and ERNIE to train the vectors, and calculated the Softmax value for the results of each model. Finally, we used LightGBM for model ensemble training.

Most existing text representation methods are based on words, phrases, sentences, or analysis of semantic and grammatical structure in texts. However, existing word segmentation techniques are not suitable in the medical field due to complex grammatical structures. Therefore, we use character-level textual representations to avoid these problems. Accordingly, our model is based on the mainstream character-level text models described below.

Figure 1. The framework of the proposed model that contains two layers: a word embedding layer consisting of 4 pretrained models (BERT, XLNet, ERNIE, and RoBERTa); and a model ensemble layer containing LightGBM, used to learn information by combining the outputs of the 4 pretrained models. BERT: Bidirectional Encoder Representations from Transformers; ERNIE: Enhanced Representation through Knowledge Integration; LightGBM: Light Gradient Boosting Machine; RoBERTa: A Robustly Optimized BERT Pretraining Approach.



BERT and RoBERTa

BERT [21] stands for Bidirectional Encoder Representation from Transformers. BERT introduces Masked Language Modeling (LM), which masks and predicts tokens in the corpus, and uses transformers [22] as an encoder to extract the contextual features of texts. The features are promoted to sentence level through sentence-level negative sampling [25], learning sentence and sentence pair representation.

Moreover, RoBERTa [26] uses dynamic masking on the basis of BERT. It removes the Next Sentence Prediction (NSP) mechanism in the pretraining process, and uses larger data for training to make RoBERTa more robust.

In this paper, we use a pretrained model based on Chinese BERT and RoBERTa with a Whole Word Masking (WWM) version [27]. In our preprocessing, a "[CLS]" symbol is added before input texts. It uses the transformer to extract features from texts and encode global information. The output of highest hidden layer at the "[CLS]" position is taken as a sentence-level feature. Subsequently, a fully connected layer is used to output text classification probability values.

Preprocessing

In natural language processing tasks, data preprocessing often greatly impacts the final result. The purpose of data preprocessing is to improve the quality of extracted text features [28]. In addition to the preprocessing carried out for different models mentioned above, we also applied some text preprocessing. First, we used regular expressions on input sentences to reduce noise characters in sentences. Subsequently, a stop word list is utilized to remove meaningless words. For sentences longer than 40 characters, we use the first 40 characters for training. To normalize input vectors, we used a dictionary to map each character to a corresponding value, and convert texts into a vector composed of numerical values.

ERNIE

Based on BERT, ERNIE [29] pretrains the model by masking semantic units such as words and entity concepts in masked LM, and enhances the semantic representation capabilities by introducing multisource data corpora.

We used a Chinese corpus–based pretrained model named ERNIE. In our preprocessing, a "[CLS]" symbol is added before input texts, and the features are extracted through a transformer with unshared weight. Here, global information is encoded into "[CLS]." Finally, we take the output of the highest hidden layer at "[CLS]" as a sentence-level feature for text classification by a fully connected layer.

XLNet

XLNet is an autoregressive language model created by Google Brain and Carnegie Mellon University, which avoids the shortcomings of the BERT model in training-tuning differences caused by using masks not existing in real texts and ignoring the relevance of cover words in prediction.

We used the XLNet [30] pretraining model based on a Chinese corpus. Text features were extracted by using Transformer-XL. The output of pooling of the highest hidden layer is used as the sentence-level feature for text classification.

Model Ensemble

In the last layer of our model, after obtaining the training output of BERT, ERNIE, XLNet, and RoBERTa, we performed Softmax processing to obtain the probability that each submodel predicts 44 labels for each text. Let the probability of each model output be  where k ∈ [1,4] represents 4 submodels, n represents the size of the training set text, and m represents that each text prediction corresponds to m different categories, set to 44 in our model. Matrix row splicing is conducted on these 4 probability variables and they are merged into a matrix of n rows and 4m columns of  as a training set. Using the idea of linear weighting, we take the training set and actual labels of n texts as input and use the LightGBM [31] to train our model to achieve the final predictions of each text.

Class Imbalance and Loss Function

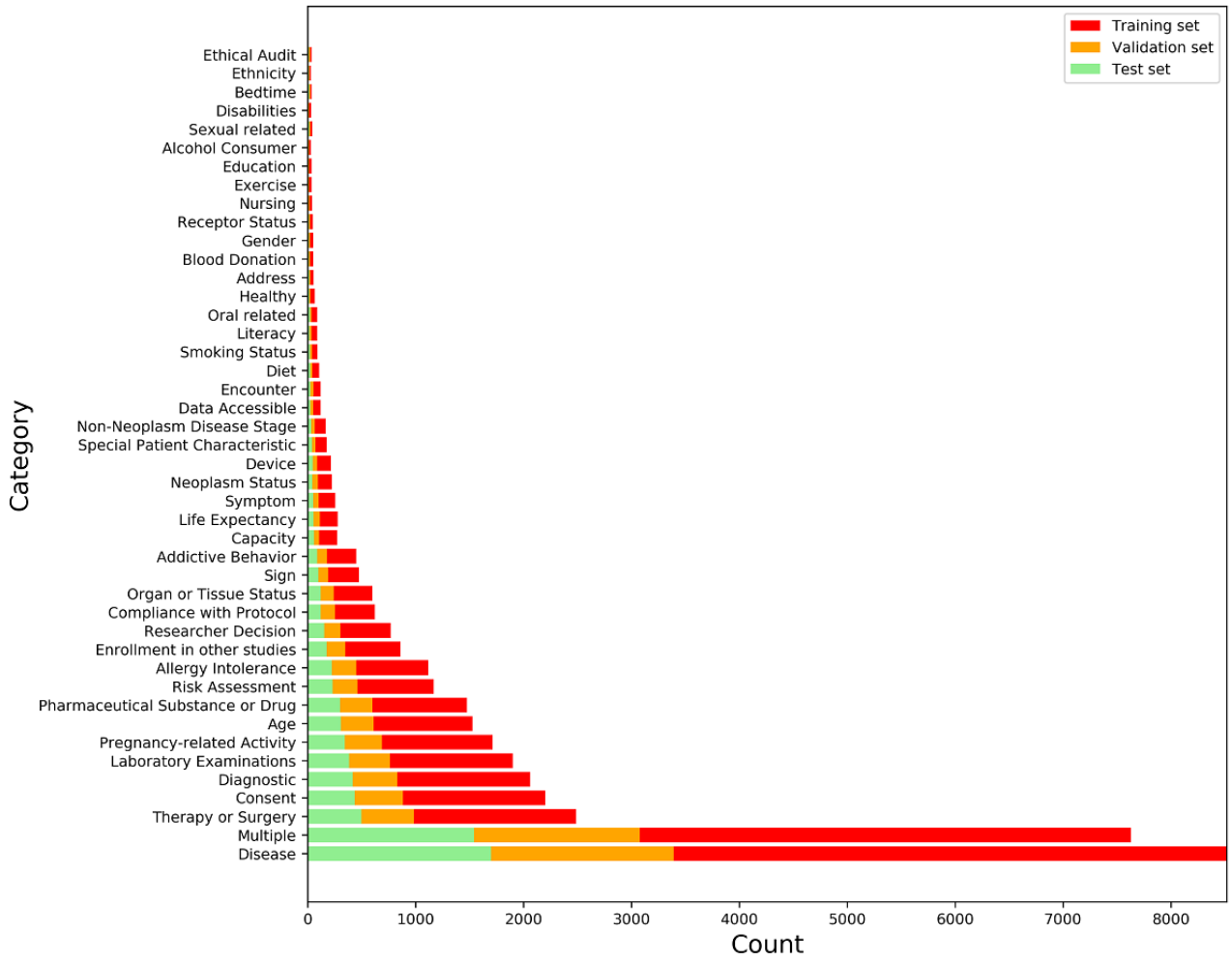
We calculated the statistical characteristics of the training, validation, and test sets, and identified that there is a data imbalance issue. Figure 2 summarizes the distributions of categories on the training, test, and validation sets to illustrate the distribution of numbers in each class. As indicated by the distribution of each data set, the data in each category is significantly unbalanced. The largest category is "Disease," with a total of 8518 samples, and the smallest category is "Ethnicity," with only 29 samples.

To solve the problem of data imbalance, we applied focal loss [32] as the loss function for training. We compared the classification of focal loss with the popular cross-entropy loss (CE loss) in the next section to show the advantage of focal loss. Supposing the expression of p_t is the following:

x_t is the score on category t, and p_t is the prediction probability of an input sample on category t. The expression of CE loss is calculated using Equation 2.

 represents the category of the i-th input sample. After that, we introduced the expression of focal loss as shown in Equation 3, where γ is a parameter. The value of γ was empirically set to 2 in this paper.

Figure 2. Histogram distributions of the training set, validation set, and test set. The y-axis represents different labels, and the x-axis represents quantity.



Experiment Setup

In order to ensure the reproducibility of test results and to facilitate the experimental comparison of different methods, this experiment fixed the random number seed to 0, the batch size was 128, and model parameters remained the same as the learning rate was set to 2×10^{-5}.

Our training used an NVIDIA 2080Ti graphics card. The memory size was 11 GB. Due to limited video memory, BERT, XLNet, and ERNIE were trained separately, including the training set (22932 pieces of data), the validation set (7652 pieces of data), and test set (7697 pieces of data). The learning rate was 2×10^{-5}, and 30 rounds of training were conducted for each model using Adam as an optimizer.

Our model was implemented using Python, based on the open source framework of PyTorch and open source pretraining parameters. To make the model converge faster and obtain better performance, we used open source parameters trained with a large amount of Chinese texts for different models for transfer learning.

Evaluation Metrics

To evaluate our model, we applied four commonly used metrics in machine learning. They are accuracy, precision, recall, and F1 score. These four metrics are also often used in classification tasks in deep learning. F1 is the standard metric for this task; it combines precision and recall. Macro F1 is a parameter index that can best reflect the effectiveness and stability of the model. According to the task requirement of CHIP 2019, we applied the macro average on these four metrics. The calculation of the four metrics is as shown in Equations 4-7:

TP (true positive) is the number of categories *t* that were correctly predicted as *t*. FP (false positive) is the number of categories that were not *t* and were wrongly predicted as *t*. FN (false negative) is the number of categories that were *t* and the model wrongly predicted it as another class. TN (true negative) is the number of categories that were not *t* and were correctly predicted as another class. In Equations 4-7, n denotes the number of categories, which is 44 in this paper.

Results

We used the current 4 single models for experiments and each model was tested on the training set only, to provide baselines for comparison. Table 2 presents the results of the 4 single models and the 2 fusion methods of Voting and LightGBM. The results from using the multimodel fusion methods were higher than that of the single models by an average of 2.35%.

By studying the loss function of the training set, we found that the performance of a single model using focal loss was significantly better using than CE loss for data sets with unbalanced categories. Figure 3 shows the convergence of loss function on the training set, in which the convergence of focal loss on the training set was faster and the value of loss function fluctuated slightly. Thus, the training speed was more stable.

Due to the structure and parameter differences of the models, the probability distributions of the models were different from each other. For a classification task, the final parameter distributions of the models were varied, and the results from different inputs had different confidences. After model assembling, a more accurate prediction result of the input sample [33] was acquired. To determine whether the performance of the classification models was limited by the amount of data, we kept the training set unchanged and randomly reduced the data volume of each category in the training set (not verification set) by 25% to keep the same data distribution. Experiments on the stability of the models were performed separately. The results are shown in Table 3.

Table 2. The performance of our model and baseline models using the full training data set.

Model	Accuracy	Precision	Recall	Macro F1
BERT[a]	0.836	0.779	0.802	0.788
XLNet	0.844	0.790	0.811	0.795
ERNIE[b]	0.836	0.786	0.795	0.783
RoBERTa[c]	0.840	0.791	0.800	0.792
Ensemble (Voting)	0.846	0.800	0.812	0.802
Our model	0.846	0.803	0.817	0.808

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]ERNIE: Enhanced Representation through Knowledge Integration.

[c]RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Figure 3. Histogram distributions of the training set, validation set, and test set. The y-axis represents different labels, and the x-axis represents quantity.

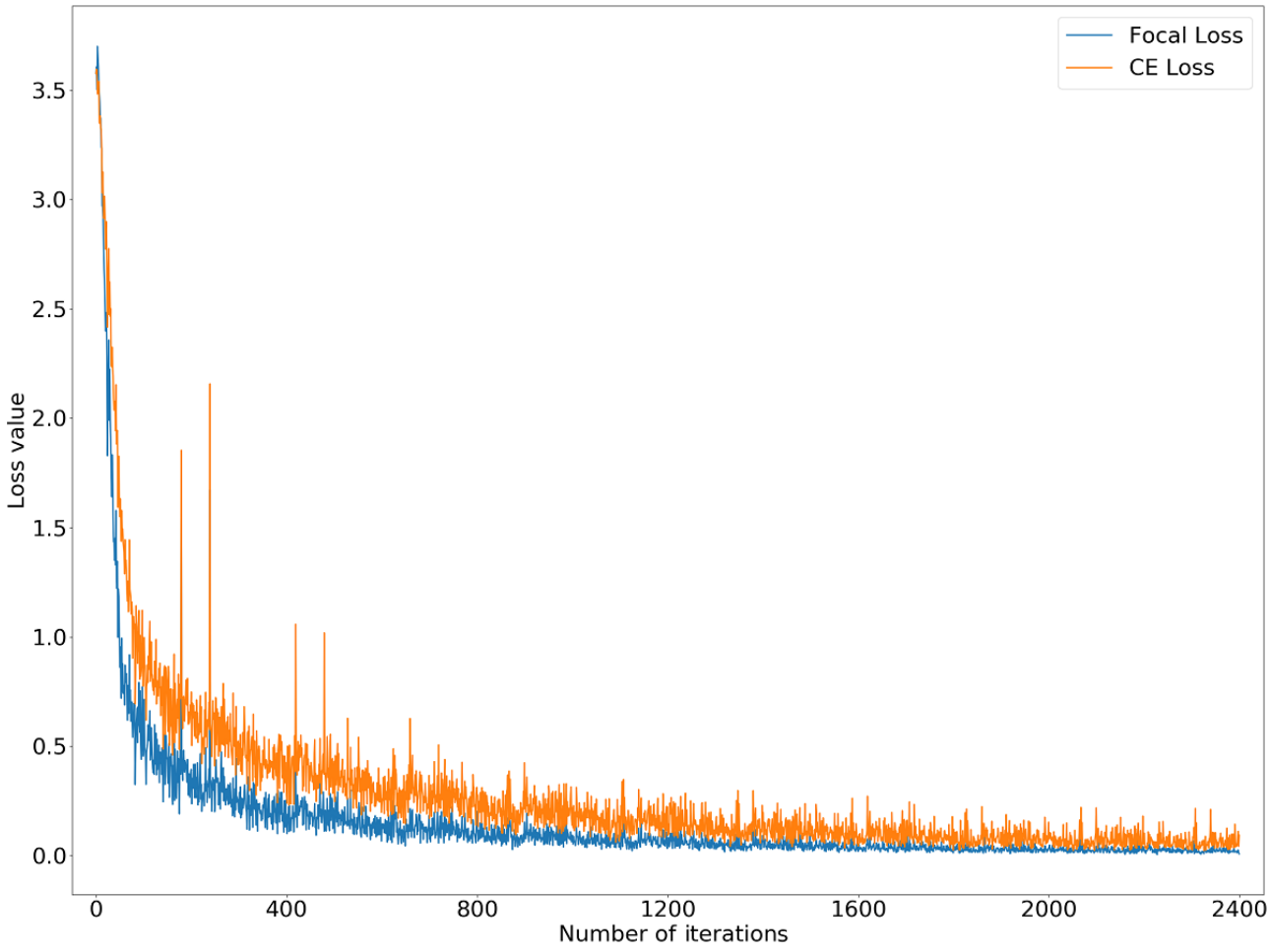


Table 3. The performance of the 6 models using the reduced training data set.

Model	Accuracy	Precision	Recall	Macro F1
BERT[a]	0.831	0.781	0.776	0.771
XLNet	0.839	0.797	0.759	0.773
ERNIE[b]	0.822	0.754	0.765	0.751
RoBERTa[c]	0.832	0.7952	0.770	0.776
Ensemble (Voting)	0.832	0.795	0.770	0.776
Our model	0.834	0.790	0.785	0.780

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]ERNIE: Enhanced Representation through Knowledge Integration.

[c]RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Discussion

Limitations

There was a limitation of the proposed method. Compared with the performance of the model under the complete data volume (Table 2), the performance of each model after reducing unequal data volume (Table 3) was significantly lower than that of the entire data volume. The F1 score of the BERT model decreased by 2.16%; the F1 score of the XLNet model decreased by 2.77%; and the F1 score of the model we proposed decreased by 3.47%. Therefore, insufficient training data is an important factor limiting model performance.

Future Work

In the future, we believe that two aspects of our model could be improved: the data and the model. Short text has the characteristic of having fewer words, and may not be able to provide enough information [34]. Therefore, a pretrained model in the medical field that was pretrained by medical corpus will benefit the stability of the model [35]. In addition, effective data enhancement could be applied on short text data to enhance text features and improve results.

Conclusions

The classification of clinical trial eligibility criteria texts is a fundamental and critical step in clinical target population recruitment. This research proposed an ensemble learning method that integrates the current cutting-edge deep learning models BERT, ERNIE, XLNet, and RoBERTa. Through model ensemble in two layers, we trained our model and compared it with a list of baseline deep learning models on a publicly available standard data set. The results demonstrated that our proposed ensemble learning method outperformed the baseline methods by 2.35% on average.

Acknowledgments

The work was supported by funding from the National Science Foundation Grant of China (U1711266), the Science and Technology Plan of Guangzhou (201804010296), and the Natural Science Foundation of Guangdong Province (2018A030310051).

Conflicts of Interest

None declared.

References

1. Zhe H, Simona C, Tianyong H, Ida S, Chunhua W. A Method for Analyzing Commonalities in Clinical Trial Target Populations. In: AMIA. 2014 Presented at: AMIA Annual Symposium 2014; November; Washington, DC, USA p. 15-19.
2. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. Journal of Biomedical Informatics 2014 Dec;52:112-120. [doi: 10.1016/j.jbi.2014.01.009]
3. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic Screening Improves Efficiency in Clinical Trial Recruitment. Journal of the American Medical Informatics Association 2009 Aug 28;16(6):869-873. [doi: 10.1197/jamia.m3119]
4. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort Required in Eligibility Screening for Clinical Trials. JOP 2012 Nov;8(6):365-370. [doi: 10.1200/jop.2012.000646]
5. Gulden C, Kirchner M, Schüttler C, Hinderer M, Kampf M, Prokosch H, et al. Extractive summarization of clinical trial descriptions. International Journal of Medical Informatics 2019 Sep;129:114-121. [doi: 10.1016/j.ijmedinf.2019.05.019]
6. Wu H, Toti G, Morley K, Ibrahim Z, Folarin AJR, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]

7. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of Predicting Health Care Utilization Via Web Search Behavior: A Data-Driven Analysis. J Med Internet Res 2016 Oct 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
8. George F. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. J Mach Learn Res 2003:1289-1305 [FREE Full text]
9. Chen IY, Agrawal M, Horng S, Sontag D. Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph. arXiv 2019 Oct 02. [doi: 10.1142/9789811215636_0003]
10. Ni Y, Zhu H, Cai P, Zhang L, Qui Z, Cao F. CliniQA: highly reliable clinical question answering system. Stud Health Technol Inform 2012;180:215-219. [Medline: 22874183]
11. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-178 [FREE Full text] [doi: 10.1136/amiajnl-2014-002887] [Medline: 25030032]
12. Huang C, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform 2016 Jan 01;17(1):132-144 [FREE Full text] [doi: 10.1093/bib/bbv024] [Medline: 25935162]
13. Li T, Zhu S, Ogihara M. Using discriminant analysis for multi-class classification: an experimental investigation. Knowl Inf Syst 2006 Mar 24;10(4):453-472. [doi: 10.1007/s10115-006-0013-y]
14. Chen B, Jin H, Yang Z, Qu Y, Weng H, Hao T. An approach for transgender population information extraction and summarization from clinical trial text. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):62 [FREE Full text] [doi: 10.1186/s12911-019-0768-1] [Medline: 30961595]
15. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
16. Kaljahi R, Foster J. Any-gram kernels for sentence classification: A sentiment analysis case study. arXiv 2017 [FREE Full text]
17. Kim Y. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October; Doha, Qatar p. 1746-1751. [doi: 10.3115/v1/d14-1181]
18. Lee JY, Dernoncourt F. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June; San Diego, CA, USA p. 515-520 URL: https://www.aclweb.org/anthology/N16-1062/ [doi: 10.18653/v1/n16-1062]
19. Hsu ST, Moon C, Jones P. A Hybrid CNN-RNN Alignment Model for Phrase-Aware Sentence Classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017 Presented at: 15th Conference of the European Chapter of the Association for Computational Linguistics; April; Valencia, Spain p. 443-449. [doi: 10.18653/v1/e17-2071]
20. Zhou P, Qi Z, Zheng S. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016 Presented at: COLING 2016, the 26th International Conference on Computational Linguistics; December; Osaka, Japan p. 3485-3495.
21. Jacob D, Ming-Wei C, Kenton L. BERT, Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/N19-1423]
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9; Long Beach, CA, USA URL: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
23. Jabs DA, Nussenblatt RB, Rosenbaum JT, Standardization of Uveitis Nomenclature (SUN) Working Group. Standardization of uveitis nomenclature for reporting clinical data. Results of the First International Workshop. Am J Ophthalmol 2005 Oct;140(3):509-516. [doi: 10.1016/j.ajo.2005.03.057] [Medline: 16196117]
24. Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D. Domain Separation Networks. 2016 Presented at: 30th Conference on Neural Information Processing Systems (NIPS 2016); December 5-10; Barcelona, Spain.
25. Kun X, Yansong F, Songfang H, Dongyan Z. Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: 2015 Conference on Empirical Methods in Natural Language Processing; September 2015; Lisbon, Portugal p. 536-540. [doi: 10.18653/v1/D15-1062]
26. Yinhan L, Myle O, Naman G, Jingfei D, Mandar J, Danqi C, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv 2019 [FREE Full text]

27. Yiming C, Wanxiang C, Ting L, Bing Q, Ziqing Y, Shijin W, et al. Pre-Training with Whole Word Masking for Chinese BERT. arXiv 2019 [FREE Full text]
28. Uysal AK, Gunal S. The impact of preprocessing on text classification. Information Processing & Management 2014 Jan;50(1):104-112. [doi: 10.1016/j.ipm.2013.08.006]
29. Yu S, Shuohuan W, Yu-Kun L, Shikun F, Xuyi C, Han Z, et al. ERNIE: Enhanced Representation through Knowledge Integration. arXiv 2019 [FREE Full text]
30. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. ArXiv 2019 [FREE Full text] [doi: 10.18653/v1/p19-1285]
31. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: NIPS 2017. 2017 Presented at: Advances in Neural Information Processing Systems 30 (NIPS 2017); Dec 4-9; Long Beach, CA, USA.
32. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. ArXiv 2017:2999-3007 [FREE Full text] [doi: 10.1109/iccv.2017.324]
33. Tebaldi C, Knutti R. The use of the multi-model ensemble in probabilistic climate projections. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2007 Jun 14;365(1857):2053-2075. [doi: 10.1098/rsta.2007.2076]
34. Wang F, Wang Z, Li Z, Wen JR. Concept-based Short Text Classification and Ranking. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 2014 Presented at: CIKM '14; November; Shanghai, China p. 1069-1078. [doi: 10.1145/2661829.2662067]
35. Radford A, Narasimhan K, Salimans T. Improving language understanding by generative pre-training. Improving language understanding by generative pre-training 2018 [FREE Full text]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers.
BILSTM: Bidirectional Long Short-Term Memory.
CHIP: China Health Information Processing Conference.
CNN: convolutional neural network.
ERNIE: Enhanced Representation through Knowledge Integration.
LightGBM: Light Gradient Boosting Machine.
LSTM: Long Short-Term Memory.
NLP: Natural Language Processing.
NSP: Next Sentence Prediction.
RNN: recurrent neural network.
RoBERTa: A Robustly Optimized BERT Pretraining Approach.
WWM: Whole Word Masking.

Edited by T Hao; submitted 15.01.20; peer-reviewed by Z Zhang, L Zhang; comments to author 14.02.20; revised version received 09.03.20; accepted 14.03.20; published 01.07.20.

Please cite as:
Zeng K, Pan Z, Xu Y, Qu Y
An Ensemble Learning Strategy for Eligibility Criteria Text Classification for Clinical Trial Recruitment: Algorithm Development and Validation
JMIR Med Inform 2020;8(7):e17832
URL: https://medinform.jmir.org/2020/7/e17832
doi:10.2196/17832
PMID:32609092

Original Paper

Medical Knowledge Graph to Enhance Fraud, Waste, and Abuse Detection on Claim Data: Model Development and Performance Evaluation

Haixia Sun[1*], MSc; Jin Xiao[2*], MSc; Wei Zhu[2], MSc; Yilong He[2], MSc; Sheng Zhang[2], MSc; Xiaowei Xu[1], MSc; Li Hou[1], PhD; Jiao Li[1*], PhD; Yuan Ni[2], PhD; Guotong Xie[2*], PhD

[1]Institute of Medical Information & Library, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China

[2]PingAn Health Technology, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Guotong Xie, PhD
PingAn Health Technology
Qianhai Complex A201, Qianwan Road 1
Qianhai Shenzhen-Hong Kong Cooperation Zone
Shenzhen
China
Phone: 86 21 38649320
Email: xieguotong@pingan.com.cn

Abstract

Background: Fraud, Waste, and Abuse (FWA) detection is a significant yet challenging problem in the health insurance industry. An essential step in FWA detection is to check whether the medication is clinically reasonable with respect to the diagnosis. Currently, human experts with sufficient medical knowledge are required to perform this task. To reduce the cost, insurance inspectors tend to build an intelligent system to detect suspicious claims with inappropriate diagnoses/medications automatically.

Objective: The aim of this study was to develop an automated method for making use of a medical knowledge graph to identify clinically suspected claims for FWA detection.

Methods: First, we identified the medical knowledge that is required to assess the clinical rationality of the claims. We then searched for data sources that contain information to build such knowledge. In this study, we focused on Chinese medical knowledge. Second, we constructed a medical knowledge graph using unstructured knowledge. We used a deep learning–based method to extract the entities and relationships from the knowledge sources and developed a multilevel similarity matching approach to conduct the entity linking. To guarantee the quality of the medical knowledge graph, we involved human experts to review the entity and relationships with lower confidence. These reviewed results could be used to further improve the machine-learning models. Finally, we developed the rules to identify the suspected claims by reasoning according to the medical knowledge graph.

Results: We collected 185,796 drug labels from the China Food and Drug Administration, 3390 types of disease information from medical textbooks (eg, symptoms, diagnosis, treatment, and prognosis), and information from 5272 examinations as the knowledge sources. The final medical knowledge graph includes 1,616,549 nodes and 5,963,444 edges. We designed three knowledge graph reasoning rules to identify three kinds of inappropriate diagnosis/medications. The experimental results showed that the medical knowledge graph helps to detect 70% of the suspected claims.

Conclusions: The medical knowledge graph–based method successfully identified suspected cases of FWA (such as fraud diagnosis, excess prescription, and irrational prescription) from the claim documents, which helped to improve the efficiency of claim processing.

(JMIR Med Inform 2020;8(7):e17653) doi:10.2196/17653

KEYWORDS

medical knowledge graph; FWA detection

Introduction

Currently, claim processing is a labor-intensive task for health insurance companies. For each claim document, the insurance inspector, who is usually a trained medical professional, needs to check whether the claim is reasonable from a clinical perspective, such as to catch any irrationality between a drug and diagnosis, or to check whether the examination is suitable for the diagnosis or symptoms. Detecting any signs of Fraud, Waste, and Abuse (FWA) is akin to looking for a needle in a haystack through claim data. The insurance company needs to hire people with sufficient medical knowledge, which significantly increases its human resource cost. Besides, claim processors still need to consult textbooks or the drug labels periodically as it is quite hard to remember details for all types of diseases, drugs, and examinations, which reduces the efficiency of claim processing. To improve the efficiency of the labor-intensive claim processing task, domain experts have devised some rules to generate a warning for suspected claims automatically. However, as the claims are coming from various hospitals that use different terminologies for drugs, examinations, and diagnoses, the coverage of fixed rules established by domain experts is relatively low. Moreover, as the drug information continues to be updated, the rules need to be updated correspondingly. To handle these challenges, knowledge graph technology could be used to represent unstructured medical knowledge such that the computer could perform reasoning on top of the knowledge graph to determine whether the claim is clinically reasonable automatically. Moreover, a method to build the knowledge graph automatically or with low human labor cost is indispensable.

Computational methods have been studied to detect FWA events [1-4]. However, it is difficult for these methods to collect comprehensive data supporting graph analysis results. Machine-learning methods failed to handle complex situations and provide interpretable evidence. There is also a gap between research and industry in FWA detection. Medical knowledge graph techniques provide a sound solution for interpretability. Recently, many medical knowledge graphs have been constructed based on medical terminology, ontology, clinical guidelines, medical encyclopedias, online forums, and electronic medical records [5-9]. For Chinese medical knowledge graph construction, natural language processing techniques have shown excellent performance on named entity recognition (NER; eg, disease, drug, and symptom) and relation extraction (eg, treatment, diagnosis) [5]. A challenge for medication information extraction from clinical notes is organization [10]. However, drug labels also contain valuable clinical information. A method that can extract high-accuracy information from drug labels is therefore expected. In addition, it is challenging to assess the effectiveness of a medical knowledge graph in an artificial intelligence app [9]. Disease-centered knowledge graphs [11,12] are tailored toward clinical decision-making support instead of using large-scale data without a curated graph. Similarly, a specific and curated medical knowledge graph is needed for enhancing FWA detection.

In this paper, we present a method to automatically build a medical knowledge graph for FWA detection in claim processing. To support FWA detection, a medical knowledge graph should cover the essential concepts such as diseases, drugs, examinations, symptoms, and the relationships between these concepts such as <*treat; drug, disease*>, <*interact; drug, drug*>, and <*check; disease, examination*>.

The main contributions of this study are as follows. First, we designed a medical knowledge graph schema for intelligent claim processing in health care insurance, and we collected recognized knowledge sources to support medical knowledge graph construction. Second, we built the medical knowledge graph using a deep learning–based method to extract entities and relationships from the knowledge sources automatically. We explored a human-machine collaboration to improve the quality of the medical knowledge graph. Finally, we applied the medical knowledge graph to empower claim processing in a health care insurance scenario.

Methods

Overview

Figure 1 shows an overview of our methodology. We divided the method into an offline workflow and online workflow. The offline workflow conducts information extraction from various medical corpora to build a comprehensive medical knowledge graph. We further improved the knowledge graph quality through domain expert review. In the online step, given the claim documents, we first identified the diagnosis and medications from the claims and then linked the mentioned terms to our medical knowledge graph. Finally, we applied the FWA rules and knowledge graph reasoning to conduct an evaluation. In the following section, we will illustrate these steps in detail.

To build the knowledge graph, we first needed to define a knowledge graph schema (ie, establish concepts and relationships) according to the requirements in claim processing. Figure 2 shows the schema of the medical knowledge graph where the circles represent the concepts and the rectangles represent the data type property. We identified three kinds of essential concepts in the FWA scenario: disease, examination, and drug. For the disease concept, as the diagnosis in the claim documents uses the International Classification of Diseases (ICD)-10 [13] terminology, we also used this terminology in the knowledge graph. For the examination, we used the terminology for the service list of China social insurance. For the drugs, we considered the Anatomical Therapeutic Chemical (ATC) level name, the generic name, and the product name. Among these concepts, we identified seven types of beneficial relationships, as shown in Figure 2 (eg, <*interaction, drug, drug*>).

Figure 1. Overview of our methodology. FWA: Fraud, Waste and Abuse.

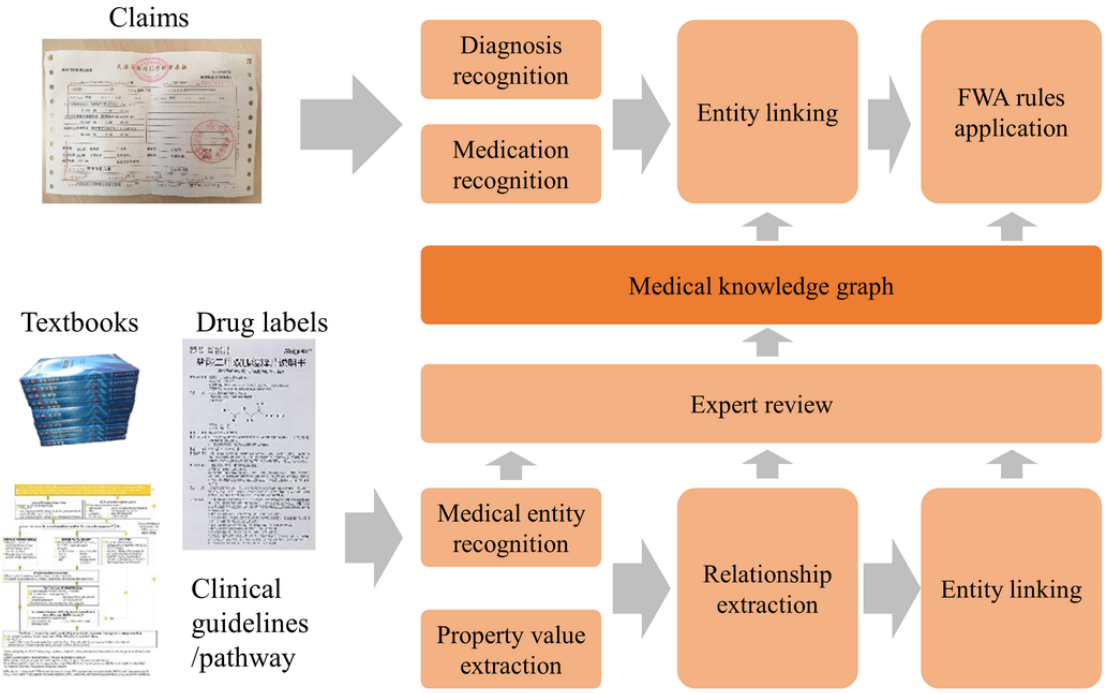
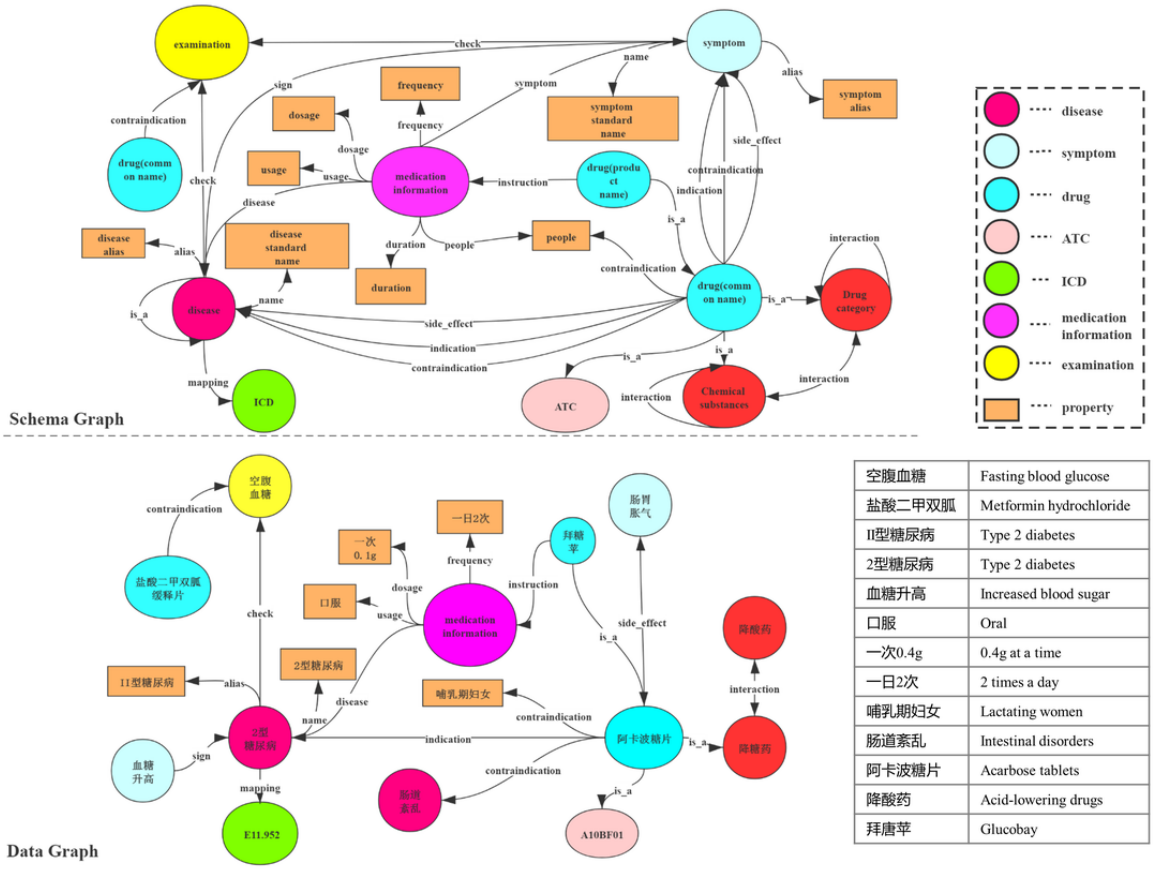


Figure 2. Medical knowledge graph schema (class) and a data graph example (instance). ATC: Anatomical Therapeutic Classification; ICD: International Classification of Diseases.



空腹血糖	Fasting blood glucose
盐酸二甲双胍	Metformin hydrochloride
II型糖尿病	Type 2 diabetes
2型糖尿病	Type 2 diabetes
血糖升高	Increased blood sugar
口服	Oral
一次0.4g	0.4g at a time
一日2次	2 times a day
哺乳期妇女	Lactating women
肠道紊乱	Intestinal disorders
阿卡波糖片	Acarbose tablets
降酸药	Acid-lowering drugs
拜唐苹	Glucobay

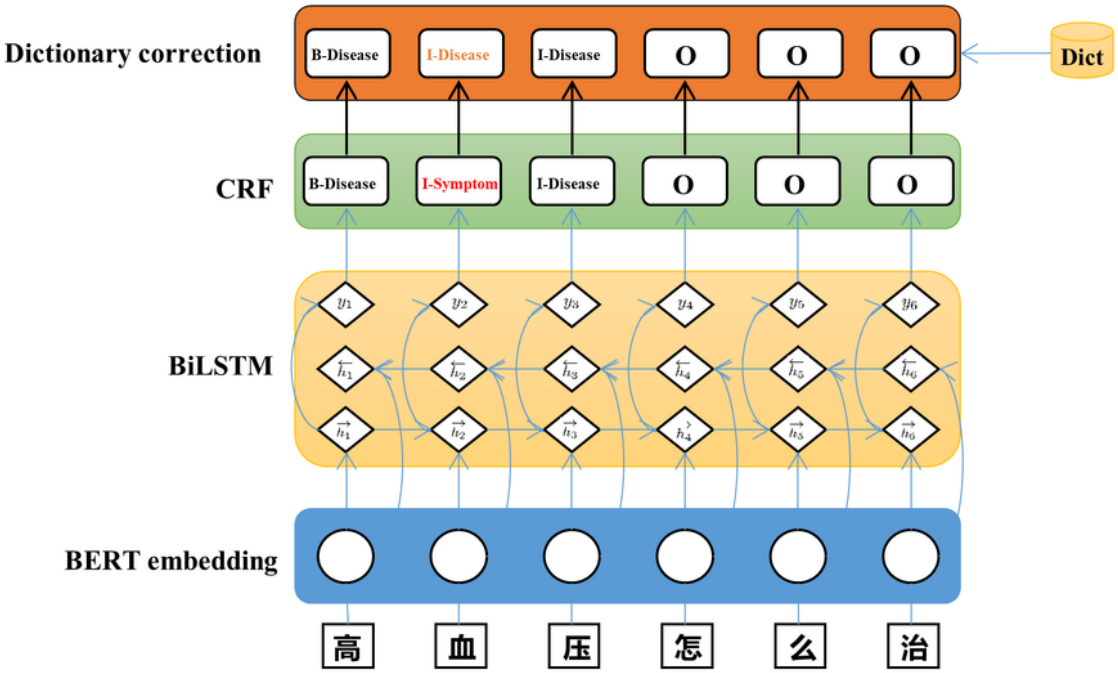
The above-required knowledge was collected from three sources: medical textbooks, drug labels, and clinical guidelines. We collected information on more than 3000 diseases and 1000 examinations from textbooks, 185,796 drug labels from the China Food and Drug Administration, and more than 2000 clinical guidelines from the Chinese Medical Association. In

the following sections, we will introduce the algorithms used to identify the concepts and relationships from these sources.

Named Entity Recognition

NER is used to detect medical entity mentions from unstructured data. As shown in Figure 3, we needed to identify five types of entities (ie, diseases, drugs, examinations, symptoms, and operation).

Figure 3. Structure of the hybrid system. BERT: Bidirectional Encoder Representations from Transformations; BiLSTM: bidirectional long short-term memory; CRF: conditional random field.



Although there are many Chinese NER methods [14-17], these methods still face many challenges, especially in the medical field. Therefore, we developed a hybrid method combining a neural network and dictionary-based system to optimize performance with limited training data, as shown in Figure 3. The input sentence first passes through the pretrained Bidirectional Encoder Representations from Transformations (BERT) model to obtain contextualized embeddings. Subsequently, there is a bidirectional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF) layer to provide preliminary predictions [18]. Finally, the predictions of the model would be corrected by a high-quality dictionary if any mistake is present. The description of the training data used is provided below.

Neural Network Model

Our model is an improved version based on conventional BiLSTM-CRF. We improved the model from the following aspects.

First, we replaced the tokenizer and word embedding with BERT [19], which is a Chinese-only pretrained language model (BERT Chinese-only model) provided by Google. By using such a pretrained language model, the effects of lacking training data can be alleviated since it provides more robust character and sentence representations.

Second, we used a feature engineering approach. We included many additional handcrafted features to the model. Neural networks have a good reputation for automatically capturing features. However, in the case of industrial application, handcrafted features can help to improve the robustness of the model. We extracted the following features. We used a word segmentation tool to extract the word segmentation soft boundary in which we used a Begin-Middle-End segmentation tag for each character of the text and the label was mapped to a lowdimensional vector by a randomly initialized matrix. Radical features were extracted, as Chinese characters are hieroglyphic, which means that the shape of each character can represent its actual meaning to some extent. In the medical domain, a character consisting of the radical “疒” is usually related to a disease or symptom. Another typical case is “月,” which is relevant to a body structure. In addition, we extracted the prefix/suffix feature. In Chinese, a word typically consists of more than one character, and some characters play the role of a prefix or suffix. For instance, a disease name often has the suffix “病” and drugs often have suffixes such as “胶囊” or “冲剂.”

Rule-Based Adjustment

Combining the predictions of a deep-learning NER model, manually developed rules, and dictionaries can be a difficult task. Results from the model and the dictionary can have conflicts, neither of which is always correct. After analyzing the results of the prediction of multiple experiments, we found that the most common mistakes that a model can make are inconsistent tagging, wrong entity type, and incomplete span. Inconsistent tagging means that a predicted entity instance is not tagged in the correct Beginning (B)-Inside (I)-Outside (O) format (eg, "I-Disease I-Disease"). A wrong entity type means the model gives out the wrong entity type. For example, it mistakes a disease for a drug, or it gives out an inconsistent entity type such as "B-Disease I-Drug I-Disease" for a disease entity. Incomplete span, the most frequently detected problem, means that the model predicts the "O" label for a part of the entity instance. For example, the model outputs a tagging sequence "B-Disease I-Disease I-Disease O O O" for the original sequence "帕金森综合症" (Parkinson disease). The above three problems can also co-occur. Thus, after we obtained the model prediction, we conducted the following adjustments. First, we checked whether the span is complete by checking whether after adding the surrounding words, the entity span is in the dictionary. If so, the longer span is accepted; otherwise, the span is accepted as is. Second, if the entity is in the dictionary, then the entity type suggested by the dictionary is used; otherwise, the entity type given by the model is accepted. If the model gives inconsistent entity types such as "BDisease IDrug IDisease," the entity type that occurs more frequently in this entity instance is chosen (ie, "Disease" in this entity instance). Finally, the tags are adjusted following the B-I-O format.

Property Value Extraction

The objective of property value extraction is to extract the property information corresponding to an entity from the unstructured text. In drug instructions, the properties we focused on mainly included the usage, frequency of administration, dosage, and treatment course, which are different when targeting different diseases or symptoms, or different populations [20,21].

Figure 4 shows an interception of the usage fields in the instructions for the metronidazole tablet. The highlighted portion indicates the properties that need to be extracted. Table 1 shows the results of the two medication information entities. The property value of the field is mostly standardized, and it is easy to summarize the template. However, the main challenge we face is that for different populations with different diseases or symptoms, the detailed usage and dosage may be different.

Figure 4. Illustration of the categories and values for properties that need to be extracted from usage section in the instruction of metronidazole tablets.

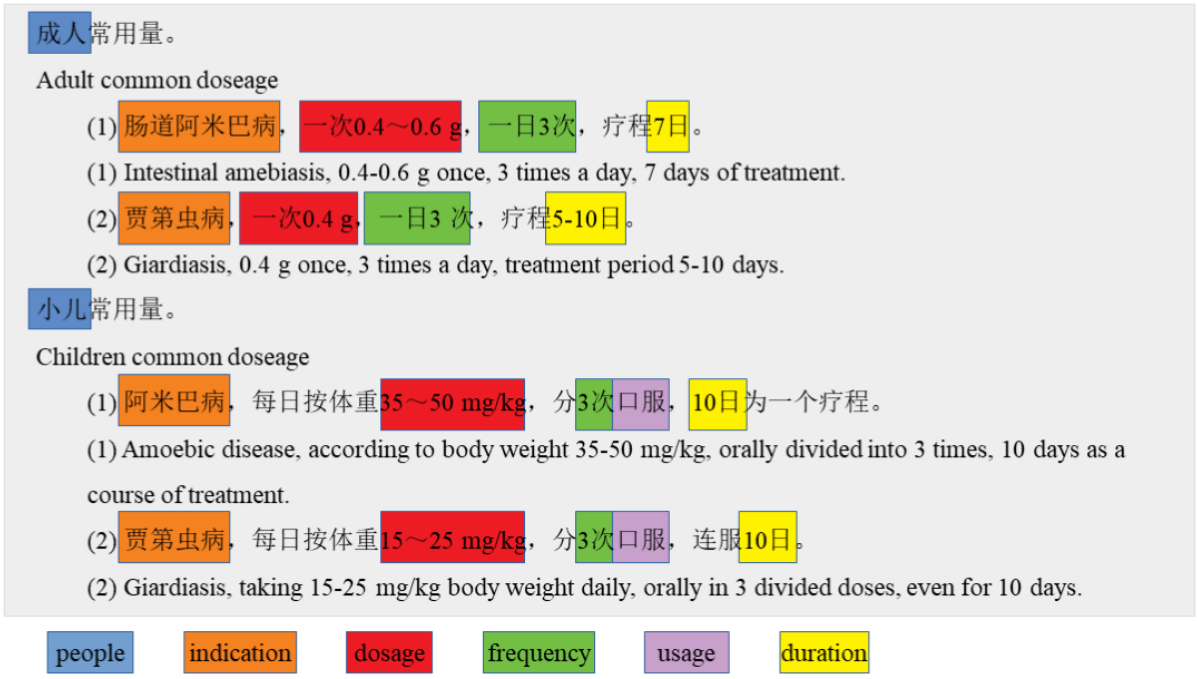


Table 1. Property values extracted from the usage section in the instructions of metronidazole tablets.

Property	Entity 1	Entity 2	Entity 3	Entity 4
Usage	口服 (Oral)	口服 (Oral)	口服 (Oral)	口服 (Oral)
Dosage	0.40.6 g/次 (g/one time)	0.4g/次 (g/one time)	3550 mg/kg	1525 mg/kg
Frequency	一日 3 次 (3 times a day)	一日 3 次 (3 times a day)	3 次 (3 times a day)	3 次 (3 times a day)
Duration	7 日 (7 days)	5-10 日 (5-10 days)	10 日 (10 days)	10 日 (10 days)
Indication	肠道阿米巴病 (Intestinal amoebiasis)	贾第虫病 (Giardiasis)	阿米巴病 (Amoebiasis)	贾第虫病 (Giardiasis)
Population	成人 (Adults)	成人 (Adults)	儿童 (Children)	儿童 (Children)

To solve these problems, property value extraction for drug instructions is usually divided into two parts: property value recognition and property value combination. Property value recognition is used to locate boundaries and determine categories, and property value combination combines property values belonging to the same entity.

Property Value Recognition

Different property values require different methods. In addition to the model-based method for indication, the remaining property values are determined by the pattern-based method [22]. The following describes the extraction method of each property value.

Dosage, Frequency, Duration, Population

The properties of dose, frequency, duration, and population are similar in form and are a combination of numbers and units; thus, similar extraction methods can be used. Taking the dose as an example, the pattern is first used to extract all combinations of numbers and dosage units such as “gram” or “slice,” and then the context keywords are used to retain the combination so that context hit keywords form the property value of dosage. The population property value is considered since a description of the taboo property may exist, such as that the dosage is 2 times a day for children but prohibited for children under 1 year of age. Therefore, interference data should be filtered according to the context keyword (prohibited) instead of the reservation.

Usage

Usage refers to the administration method of a drug such as “口服” (oral). We developed a set of patterns, which are shown in Table 2, to extract usage property values. For drug instructions that do not specify usage, we built a mapping table to infer it according to the dosage form, as shown in Table 3.

Table 2. Patterns of property value extraction described by regular expression-like syntax.

Pattern	Description	Example
一[次日](num)+片	Dosage pattern; [num] refers to an integer or decimal	一次一粒 (One capsule at a time)
(num)+[片粒克…](/kg)	Dosage pattern for drugs by weight	10毫克/kg体重 (10 mg/kg body weight)
一日(0-9)+次	Frequency pattern for the frequency of medication by day	一日3次 (t.i.d[a])
疗程(0-9)+[日天]	Duration pattern	连续10天 (10 consecutive days)
(0-9)+岁(以[上下])?	Population pattern for age	18岁以下患者 (Patients under the age of 18)
[口嚼吞泡饮]服	Usage pattern for oral drugs	饭后口服 (Orally after meals)
静脉.{0,2}注射	Usage pattern for intravenous injection	静脉内注射 (intravenous injection) 140 mg/ml

[a]t.i.d: ter in die (3 times a day).

Table 3. Mapping table of the dosage form to usage.

Dosage form	Usage
缓释片(sustained release tablets)	口服 (oral)
肠溶片 (enteric-coated tablets)	口服 (oral)
直肠栓剂 (rectal suppositories)	肛门用药 (anal medication)
贴剂 (patch)	外用 (external)
洗剂 (lotion)	外用 (external)

Indication

An “indication” for a drug refers to the use of that drug for treating a disease. For example, diabetes is an indication for insulin. Following the previous section, we employed a BERT-BiLSTM-CRF model with manually designed features for indication detection, and the detected entities were linked to our medical knowledge graph.

Property Combination

The property values of the same medication information entity need to be combined. In most cases, all types of property values for the same entity will appear in the same sentence. For a small number of the remaining cases, we aggregated the extracted information via three heuristic rules based on linguistic patterns in the drug instructions as follows: (1) usage, dosage, and duration usually appear at the end; (2) if a property value does not appear in the description of the current entity but appears in the previous entity, the property value is usually the same as the previous entity; (3) if the population changes, the disease, frequency, and duration will also change.

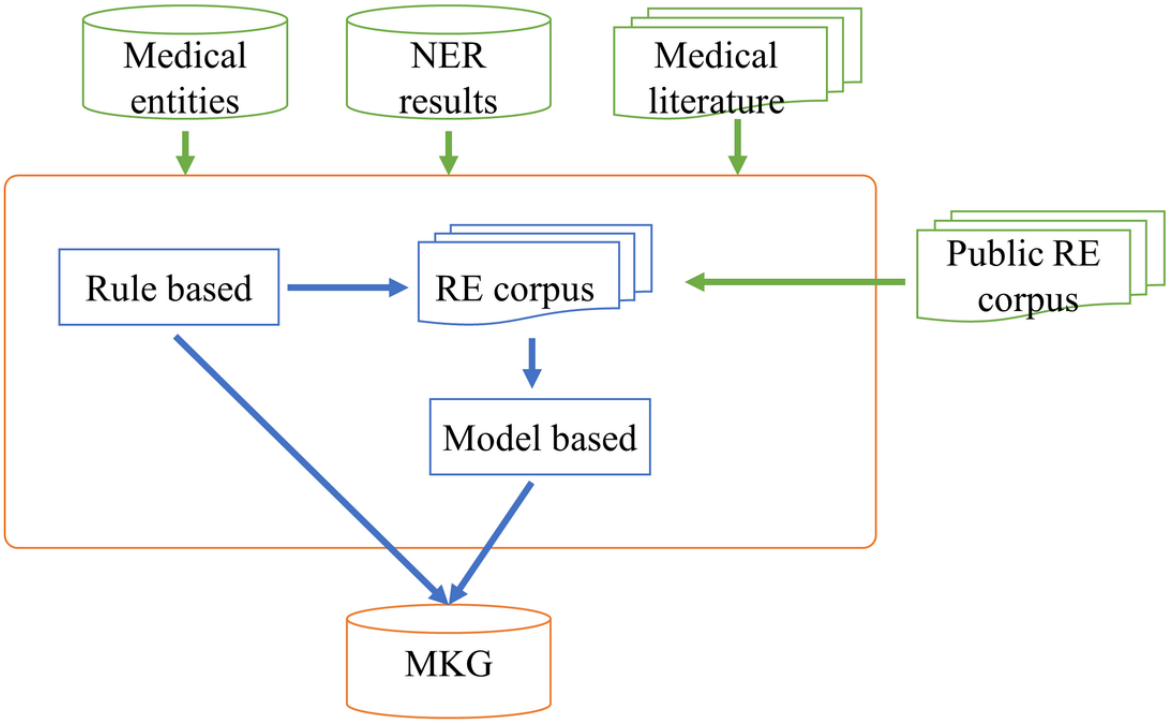
Figure 4 and Table 1 illustrate the above three rules. Following rule (1), the duration (eg, “10日为一个疗程”, 10 days as a treatment course) appears last in the sentence. For entity 4, the sentence mentioning it does not specify the population group. However, following rule (2), we know that the population should be children, according to entity 3. All properties of entity 2 and entity 3 are different because of the change of population, as indicated by rule (3).

Therefore, most of the property values in the same sentence can be directly combined. Otherwise, we manually combined the extracted information to ensure precision and improve performance.

Relation Extraction

Medical relation extraction refers to the semantic relationship between medical entities defined in the medical knowledge graph schema [23]. The main types of medical relations considered in this study include drug-drug interactions (DDIs), indications, and contraindications. Detailed information of the dataset is discussed further below. The medical relationship extraction framework proposed in this paper mainly includes two parts: a distant supervision method and a model-based method, as shown in Figure 5. In the distant supervision method, the medical relation extraction templates are formulated based on partofspeech, syntactic structure, specific keywords, and expert medical knowledge [24]. The precision of medical relationships extracted by rule-based methods is high, but the recall is low. The amount of relations acquired by the rule-based method depends on the quantity and quality of the templates. In the model-based method, deep-learning models, especially those employing an attention mechanism, automatically contextualize the entity pairs together with their context information. Thus, they can generalize well and improve the recall of relation extraction [25].

Figure 5. Framework of medical relation extraction (RE). MKG: medical knowledge graph; NER: named entity recognition.



Rule-Based Relation Extraction

We use the DDI relation as an example to demonstrate how to use rules for relation extraction. The relationship between drugs defined in the schema of the medical knowledge graph is divided into three categories: promotion, contraindication, and none (no relationship). Promotion indicates that two drugs can promote the efficacy of each other, contraindication means that two drugs will cause adverse reactions when taken at the same time, and none is no interaction between the two drugs. Some representative examples and patterns are provided in Table 4 and Table 5. The patterns were summarized manually after reading a small portion of the data. After the text data passes through an NER model, the entity instances are replaced with entity type symbols (eg, "吲哚美辛与胰岛素一起使用可以加强降糖效果," indomethacin, when used with insulin, can enhance the hypoglycemic effect, becomes "[Drug]与[Drug]一起使用可以加强降糖效果," [Drug], when used with [Drug], can enhance the hypoglycemic effect), and the patterns can identify the relation between the two entities.

Table 4. Categories of drug-drug interaction relations.

Categories	Explanation	Example
Promotion	Promote the efficacy of each other	吲哚美辛与胰岛素一起使用，可以加强降糖效果 (Indomethacin, when used with insulin, can enhance the hypoglycemic effect)
Contraindication	Produce adverse reactions when taken at the same time	吲哚美辛与秋水仙碱合用时可增加长胃溃疡及出血的危险 (Indomethacin combined with colchicine can increase the risk of long stomach ulcers and bleeding)
None	No interaction between the two drugs	阿德福韦酯和拉米夫定合用，两种药物的药代动力学特征都不改变 (When adefovir dipivoxil and lamivudine are used together, the pharmacokinetic characteristics of both drugs remained unchanged)

Table 5. Patterns for extracting drug-drug interaction relations.

Pattern	Description	Example
[…]禁止[…]合用	Contraindication; the sentence contains “禁忌” and “合用”	硝苯地平禁止与利福平合用 (Nifedipine is prohibited to be used with rifampicin)
[…]和[…]配伍禁忌	Contraindication; the sentence contains “配伍禁忌”	氯甲苯酸与青霉素有配伍禁忌 (Chlorotoluenic acid is contraindicated with penicillin)
[…]干扰[…]作用	Contraindication; the sentence contains “干扰” and “作用”	磺胺嘧啶片有可能干扰青霉素类药物的杀菌作用 (Sulfadiazine tablets may interfere with the bactericidal action of penicillin drugs)
[…]加强[…]效果	Promotion; the sentence contains “加强” and “效果”	盐酸昂丹司琼口腔崩解片与地塞米松合用可加强止吐效果 (Ondansetron hydrochloride orally disintegrating tablets combined with dexamethasone can enhance the antiemetic effect)
[…]增强[…]疗效	Promotion; the sentence contains “增强” and “疗效”	维生素C可增强盐酸吗啉胍注射液的疗效 (Vitamin C enhances the efficacy of morpholine hydrochloride injection)
[…]与[…]无相互作用	None; the sentence contains “无相互作用”	扎来普隆胶囊与帕罗西丁无相互作用 (Zaleplon capsules have no interaction with paroxetine)
[…]不改变[…]作用	None; the sentence contains “不改变” and “作用”	苯磺酸氨氯地平片不改变华法林的凝血酶原作用时间 (Amlodipine besylate tablets did not change the prothrombin time of warfarin)

Model-Based Relation Extraction

We experimented with a series of models for our relation extraction tasks, including piecewise convolutional neural networks (PCNN) [26], BiLSTM [18], and PCNN with adversarial training (PCNN+AT) [27]. Finally, a comparison was made among the models.

Figure 6 depicts the PCNN model [26]. The sentence is first transformed into vectors. A convolution kernel is then applied, followed by a piecewise max pooling operation. Finally, the pooled features are sent to a softmax classifier to predict the relationship between two entities. To further improve the robustness of the model, we applied AT to improve the robustness of classifiers to small worst-case perturbations by calculating the gradient direction of loss function to the data. Since AT generates continuous disturbances, we added antagonistic noise at the word-embedding level. The network is shown in Figure 7.

Figure 6. Piecewise convolutional neural networks architecture.

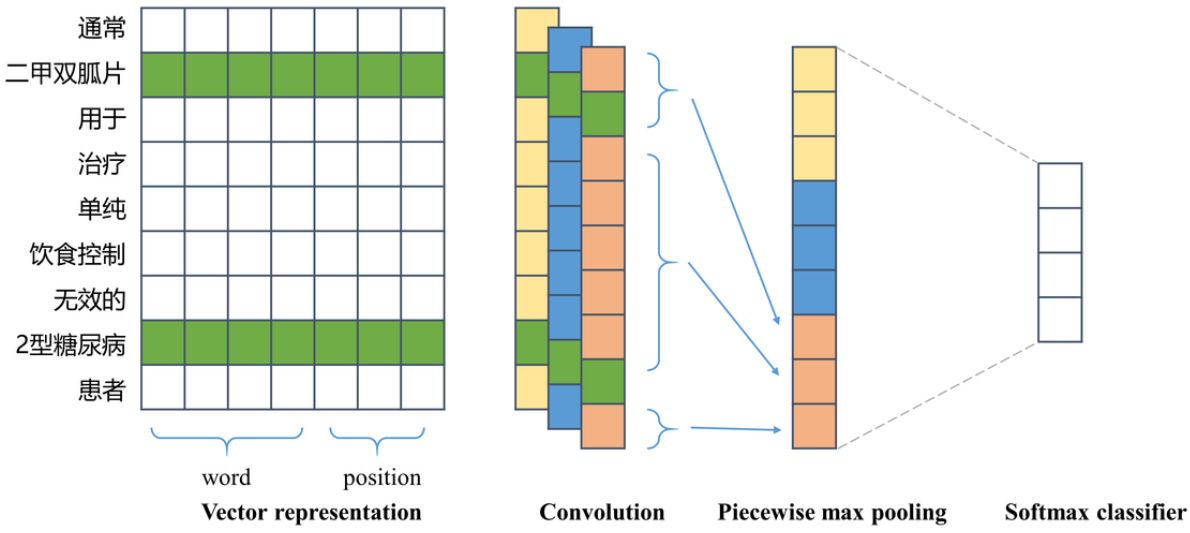
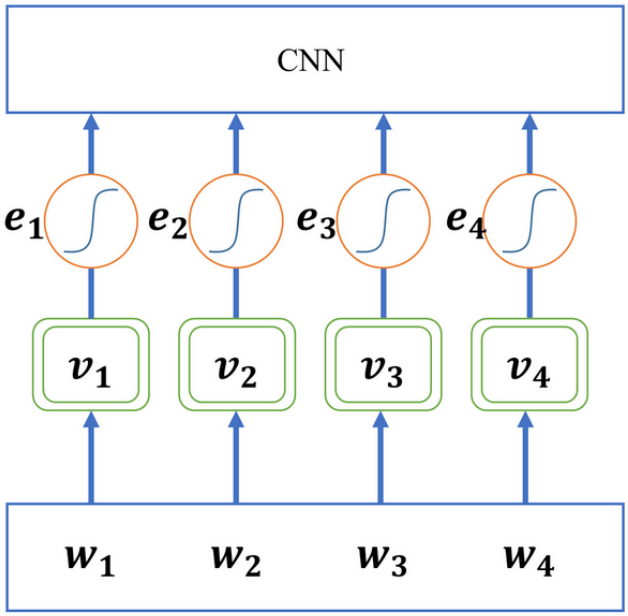


Figure 7. Computation graph of encoding a sentence xi with adversarial training. ei denotes the adversarial perturbation xi. Dropout is placed on the output of the variables in the double-lined rectangles. CNN: convolutional neural network.



Knowledge Graph Fusion

Knowledge graph fusion can be regarded as an ontology alignment task in our workflow, which has been studied extensively in the literature [28-30]. In this paper, we present the task with unique domain characteristics in the medical field on fusing knowledge cards.

The previous entity extraction step would introduce the entity mentions that are unknown terms in the existing medical knowledge graph. In this case, one must decide whether an entity mention is a variant of some term in the medical knowledge graph or a new entity, which requires a precise entity normalization system. To build such a system, three difficulties are encountered. First, typos or discrepancies in transliterations may occur in online documents; for example, “唐尿病” is a frequent typo of “糖尿病” (diabetes), akin to diabites (misspelling) and diabetes, and “咪康唑” (miconazole) and “密康唑” (miconazole) are both transliterations of miconazole. Second, some entity mentions look quite alike, but represent quite distinct entities; for example, “ii 型糖尿病” (type 2 diabetes) and “i 型糖尿病” (type 1 diabetes) are quite similar, but they are very different entities. Third, there can be some discrepancies in expressing the same component of an entity name; for example, “手部擦伤” (hand abrasion) can be easily expressed as “手擦伤” (hand abrasion) since “手部” and “手” both mean hand.

The above three difficulties make it nontrivial to build a term canonicalization system, and previous systems have not addressed the above issues altogether [31-33]. One might consider combining more sophisticated machine-learning models such as neural networks along with the lexicon features and edit distances. However, sophisticated machine-learning models are challenging to train with limited labeled data and their results are not explainable.

To effectively address the above difficulties, we designed a multilevel fine-grained similarity score system. First, on the whole, we built a multilevel string matching algorithm, which we call ZhFuzzyED, considering three levels (token, radical, and pronunciation edit distance) so that the similarity score is less sensitive to typos and transliteration differences. Second, diving deeper into the components of entity names, we found that an entity mention such as a disease entity mention usually consists of semantic units such as body structure, negation, degree adverb, some adjective describing the type or stage, and the core term that defines the disease (as shown in Figure 8). Based on this observation, we collected and categorized 11 groups of semantic units. For each semantic unit category, a subgraph can be built to measure the similarity score between two semantic units. For example, “手臂” (arm) and “前肢” (forelimb) are similar, although their surface forms are different. Similarity scores concerning different semantic unit categories are weighted along with the string level similarity score.

Natural language processing in the medical field is complicated and challenging. Thus, models and algorithms sometimes fail, and manual verification is essential. Therefore, we developed a tool (web app) to enable human-machine cooperation for knowledge graph fusion and knowledge graph quality control. The main design of the app is shown in Figure 9.

Figure 8. Sample of semantic unit's category in disease.

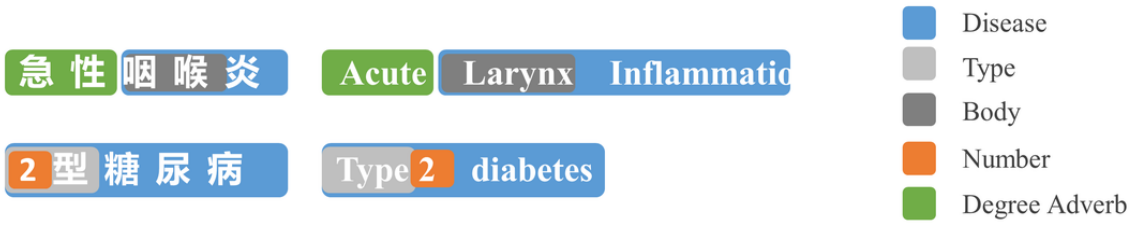


Figure 9. Design of the knowledge correction system. We examined the information extracted automatically, corrected the errors, and included the information in the medical knowledge graph, which was made available for the next round of information extraction and downstream tasks.

Concept	Waiting correct	Relation	Operation
Type 2 diabetes	Metformin sustained release tablets	indication	pass correct wait delete
Type 2 diabetes	Diabetes	is_a	pass correct wait delete
Type 2 diabetes	Hydrocortisone	contraindication	pass correct wait delete

In this step, when a new entity mention comes in, we first search in the medical knowledge graph (usually via an invert index) for a possible matched entity and then the candidates are reranked via the above similarity score system. If the best-scored candidate still obtains a low score, the entity mention is considered to be an unknown entity, waiting to be added or corrected manually by experts. Otherwise, it is considered a term for best-scored known entity. This process is equivalent to a cycle of a selflearning process since the new terms added to the medical knowledge graph can improve the accuracy of our workflow at the next round of iteration.

Applying the Medical Knowledge Graph for Claim Processing

In this section, we discuss how to use the medical knowledge graph to conduct automatic FWA detection in claim processing. Given a claim document, in the first step, we need to identify the diagnosis, examinations, and medications in the claims.

As the medical entities in a claim are extracted by optical character recognition or from various hospital information systems, these terms may follow different terminologies and may contain errors. Thus, term normalization is the foundation. We first used the aforementioned multilevel string matching algorithm (ZhFuzzyED) to perform term normalization.

After the entity mentions in a claim were linked to entities in the medical knowledge graph, we checked the following three suspicious scenarios.

Fraud Diagnosis

Fraud diagnosis is suspected when the disease does not match the indication of treatment. In this condition, the relation between a drug and disease can be used for detecting the mismatch case. There are three types of scenarios: (1) a drug does not have the disease as an indication; (2) the disease is a contraindication of the drug; and (3) no suitable drugs for treating the disease appear in this claim.

Excess Prescription

Excess prescription refers to excessive medical care such as one disease corresponding to many drugs in a claim, which is not medically necessary.

Irrational Prescription

Drugs prescribed in a claim have interactions. If the drugs in a visit record have interactions, especially when the interaction is harmful, the claim is considered to be fraudulent.

Inferring new facts from existing knowledge graphs is a form of an explainable reasoning process. The above scenarios could not be directly queried from the medical knowledge graph. Therefore, further reasoning on queries is required. Multihop knowledge graph reasoning was applied for our FWA detection. The graph reasoning rules are shown in Table 6.

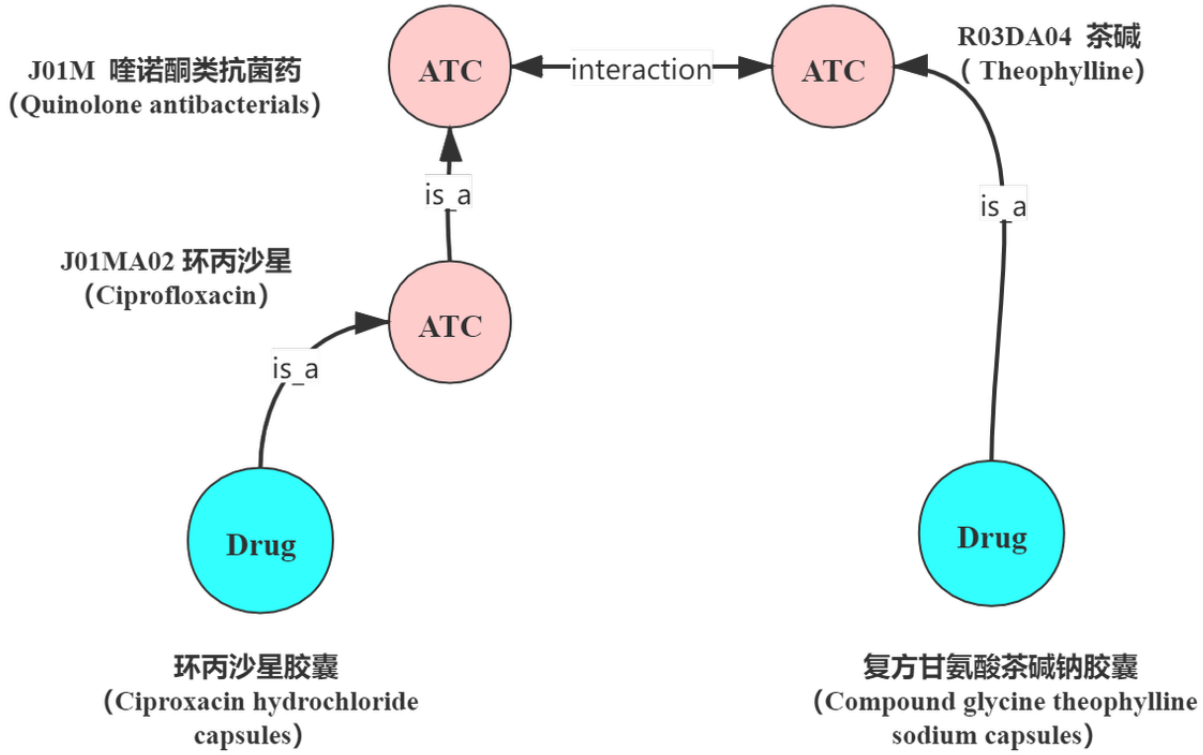
For example, as shown in Figure 10, the drug interaction relations are applied on the ATC level, whereas the relations are usually derived from the level of the common drug name (generic chemical name of a drug) extracted from the drug instructions. The occurrence of drug interactions is usually due to the chemical composition of a drug, which is the ATC code. Thus, to check whether two drugs have an interaction can provide an extension of query on ATC concepts.

Table 6. Graph reasoning processing in Fraud, Waste, and Abuse detection.

Suspicious scenarios	Graph reasoning rule (Cypher-like syntax)
Fraud diagnosis	(Disease)-[:is_a]-(Disease)<-[:indication]-(Drug)-[:is_a*0..1]->(ATC[a])<-[:is_a*0..1]-(Drug)
Excess prescription	(Drug)-[:is_a]->(ATC)<-[:is_a]-(Drug)-[:indication]->(Disease)-[:is_a]-(Disease)
Irrational prescription	(Drug)-[:is_a]->(ATC)-[:is_a*0..3]->(ATC)<-[:is_a*0..3]-(ATC)<-[:is_a]-(Drug)

[a]ATC: Anatomical Therapeutic Classification.

Figure 10. Example of graph query and graph reasoning. ATC: Anatomical Therapeutic Classification.



Results

Datasets for Model Training

Our NER corpus was drawn from drug descriptions, encyclopedia pages for medical entities, and the literature so that the model trained can adapt to different scenarios. We prioritized documents that are related to entities that are common or medically important, which were split into 10,889 sentences. The annotation process followed the majority voting rule; that is, if two annotators did not agree on the annotation of the same sentence, then a senior annotator, who is a more experienced medical practitioner, made the final annotation. To save labor costs, our annotation is carried out in an active learning fashion as introduced by Chen et al [34]. For example, we first annotated the first 500 sentences using a medical dictionary, and then annotated them fully. Following the uncertainty-based sampling method, a pool of 1500 sentences was sampled. After the 2000 sentences were annotated, a better model could be obtained on the larger dataset. After repeating this step for a few iterations, we obtained our annotated dataset with less labor and higher quality in the sense that the models trained on it will perform better than a random sampled dataset.

The relation extraction dataset was built on the same corpus. The preannotation takes advantage of the technique of distant supervision in addition to active learning [35]. Distant supervision means that if two entities in a sentence are both in the medical knowledge graph, we assume that their relation in the sentence is in agreement with their relation in the medical knowledge graph. In the active learning procedure, if distant supervision detects relations in a sentence, we will prioritize on annotating this sentence. Annotators are responsible for determining that the distant supervised relation instance is correct, and whether there are other relation instances in the sentence. The whole annotation procedure gives out 21,657 relation instances, and the labor cost is estimated to be reduced by 4.3 times due to distant supervision and active learning.

Model Performances

Performance of Named Entity Recognition

The annotated dataset was split into 8:2 training:test datasets. We compared three kinds of NER models: the deep-learning model only, the model with hand-crafted features, and the model with hand-crafted features and manually designed rules. Detailed results on the test set are shown in Table 7, demonstrating that the hand-crafted features are effective for performance

improvement. In addition, the designed rules could further improve the performance of NER significantly.

Table 7. Performance of the named entity recognition models based on the entity level F1 value.

Variable	Model only	Model+feature	Model+feature +rules
Disease	0.901	0.921	*0.924* [a]
Symptom	0.792	0.793	*0.801*
Examination	0.742	*0.744*	0.742
Drug	0.798	0.806	*0.821*
Operation	0.763	0.772	*0.784*
Overall	0.833	0.841	*0.850*

[a]Values in italics indicate the best performance on the same dataset.

Relation Extraction

The annotated dataset was split into 8:2 training:test sets. For relation extraction, we conducted experiments to evaluate the effectiveness of the three models and report the performances on the test set in Table 8. The PCNN and PCNN+AT models were described in the Methods section. The convolutional neural network (CNN) model is simply the PCNN model with vanilla pooling instead of piecewise pooling. We observed that the piecewise pooling is import for adequately representing the features of a sentence in the relation extraction task. Moreover, the PCNN+AT model had the best performance since it is more robust.

Table 8. Results of each model for overall relation extraction.

Model	Precision	Recall	F1 score
PCNN[a]	0.68	0.80	0.73
CNN[b]	0.55	0.76	0.64
PCNN+AT[c]	*0.75* [d]	*0.84*	*0.80*

[a]PCNN: piecewise convolutional neural network.

[b]CNN: convolutional neural networks

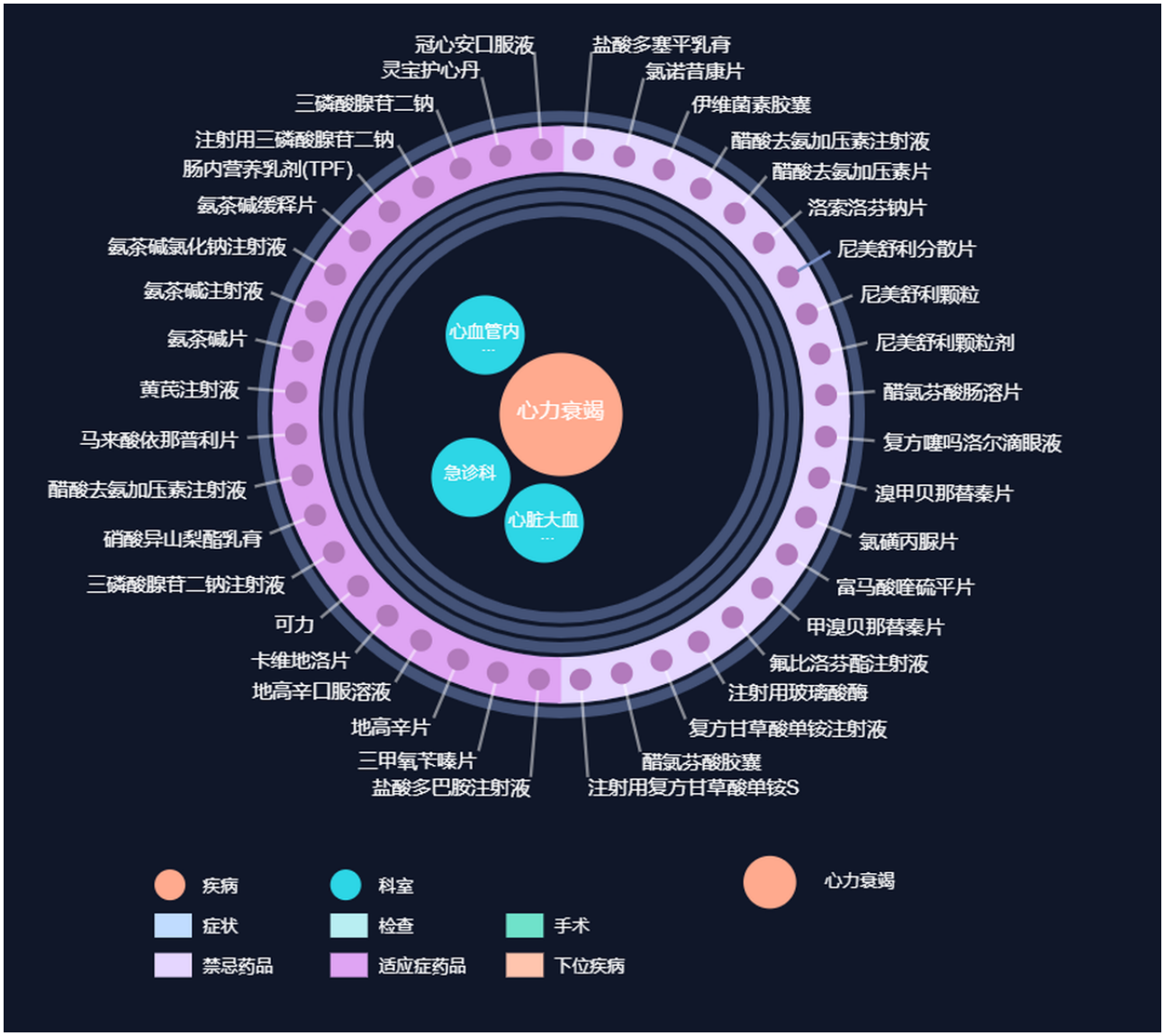
[c]AT: adversarial training.

[d]Values in italics indicate the best performance.

Statistics of the Medical Knowledge Graph

Finally, we built a medical knowledge graph that includes 1,616,549 nodes and 5,963,444 edges. To make it easier to explore the graph, we developed a web app to support the browsing on our medical knowledge graph on a website [36], which is open and free to access. Figure 11 shows a snapshot of our knowledge graph data. In brief, when the user selects a concept, the concept will be shown in the center of the circle. The concepts belonging to the same category that show certain relationships with the central concept will be placed on the same ring, whereas different types of relations will have different colors for concepts on the same ring. For example, as shown in Figure 11, the node “心力衰竭 (heart failure)” is in the center. All drugs that are related to heart failure are on the same ring. The drugs having indication relations are in dark purple while those having contraindication relations are in light purple.

Figure 11. Snapshots of our knowledge graph data.



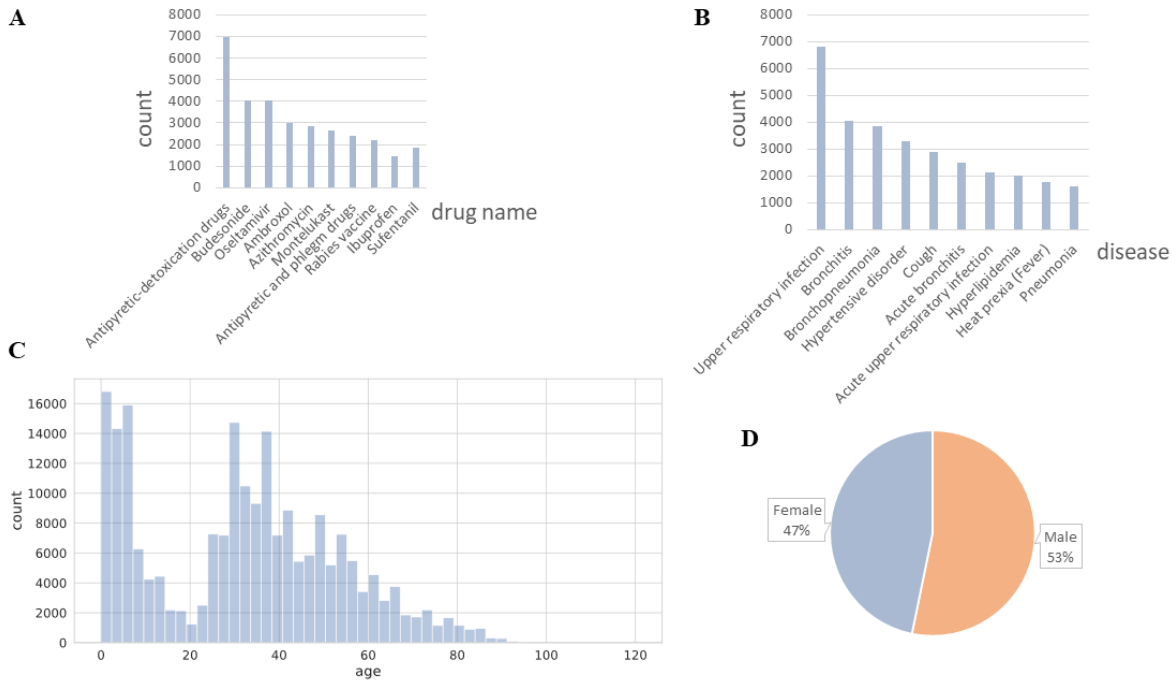
FWA Detection in Claim Processing

Dataset

We collaborated with the insurance company in our PingAn group and obtained 214,107 claim documents. Every claim document contains a list of diagnoses (1.5 diagnoses on average) and a list of drugs (2.3 drugs on average). There are 2586 unique ICD-10 codes and 5307 unique common drug names in these claim documents. More information is shown in Figure 12.

In the following, we report the performance of each step in the FWA detection process.

Figure 12. Statistical overview of claim documents: (A) top 10 drug names in claim documents; (B) top 10 diseases occurring in claim documents; (C) age distribution in claim documents; (D) sex ratio in claim documents.



Subtask: Terminology Standardizing

As described above, the first step in FWA detection is to link the diagnosis and medications to the entities in our medical knowledge graph. Our proposed multilevel string matching algorithm ZhFuzzyED achieved 0.861 accuracy in linking the diagnosis to the ICD-10 coding system and 0.902 accuracy for drug normalization.

Subtask: Graph Reasoning–Based Relation Detection

For claim processing, 10% of claims are typically rejected for various reasons. The clinical unreasonable problem is only one of the reasons for rejection. We randomly selected 100 rejected claim documents and let the insurance inspector manually label the type of the rejected reasons. We then applied our proposed FWA detection method to identify the three types of frauds as described above. Table 9 lists the number of events that were labeled by humans and the number of events that were detected by the medical knowledge database–based method. Specifically, excess prescription means the drug has been abused in a document, fraud diagnosis reflected that there is no drug suite for the diagnosis, and irrational prescription is when a conflict exists in the drug list.

Our method could help detect around 70% of these events. This result is much better than the existing method that relies on only a human to check part of the claims randomly. Therefore, the existing method requires investing many professionals and spending a substantial amount of time to check each claim one by one.

Table 9. Performance of claim processing.

Subtask	Events (n)	Detected by the MKG[a] (n)	Graph reasoning rule (Cypher-like syntax)	Claim example
Excess prescription	7	5	(Disease)-[:is_a]-(Disease)<-[:indication]-(Drug)-[:is_a*0..1]->(ATC)<-[:is_a*0..1]-(Drug)	Diagnosis: Acute bronchitis Drug: Bifidobacterium double live bacteria powder (*excess prescription*)
Fraud diagnosis	11	7	(Drug)-[:is_a]->(ATC)<-[:is_a]-(Drug)-[:indication]->(Disease)-[:is_a]-(Disease)	Diagnosis: Guillain-Barré syndrome, *Thyroid nodules (fraud diagnosis)* Drug: Immunoglobulin injection
Irrational prescription	4	3	(Drug)-[:is_a]->(ATC)-[:is_a*0..3]->(ATC)<-[:is_a*0..3]-(ATC)<-[:is_a]-(Drug)	Drug: *Atorvastatin calcium tablets, Ketoconazole cream (there is an interaction between the two drugs)*

[a]MKG: medical knowledge graph.

Discussion

Principal Results

In this paper, we have proposed an automatic method to extract information from medical knowledge to build a medical knowledge graph specifically for FWA detection. First, our NER results showed that by integrating the hand-crafted features with the embeddings helps to improve the accuracy of medical entity recognition. In addition, when the domain-specific rules were added, the performance could be further improved as shown in Table 7.

Second, for medical relation extraction, the PCNN+AT model showed better performance as compared to CNN or PCNN. Third, we constructed a high-quality medical knowledge graph, including 1,616,549 nodes and 5,963,444 edges. Finally, we designed the rules on top of the medical knowledge graph to detect three kinds of FWAs in claim processing. The experimental results showed that our approach helps to detect 70% of these FWA events automatically. The medical knowledge graph–based method provided good interpretability of the results. The reasoning process on the medical knowledge graph can help the insurance inspector to quickly determine whether the claim should be rejected, which will contribute to substantial savings in the claim processing cost. Our system has already been deployed as a service to generate alerts of suspected claims for insurance inspectors within our PingAn group.

Limitations

Our medical knowledge graph and proposed rules could detect three kinds of FWA issues. However, there are still other types of FWA events such as medication overdose and medications that are not suitable for the population. Therefore, we need to integrate more information into our medical knowledge graph and design more rules to detect more types of FWA problems. In addition, our method still missed some FWA events. This is because we failed to extract some drug categories from the drug labels. Therefore, we need to further improve the recall of our information extraction method.

Conclusions

In this study, we examined the effectiveness of building a medical knowledge graph to enhance FWA detection on claim data. Our method can help insurance inspectors to identify insurance claims worthy of attention from thousands of documents and ultimately reduce Medicare and Medicaid spending.

Acknowledgments

This work is supported by the National Key R&D Plan of China (grant no. 2016YFC0901901), Fundamental Research Funds for the Central Universities (grant nos. 2018PT33024 and 2017PT63010), and PingAn Health Technology (grant no. PMS201836894755-1/1).

Conflicts of Interest

None declared.

References

1. Umair A, Aftab A, Mohammad JS. Knowledge Representation and Knowledge Editor of a Medical Claim Processing System. J Basic Appl Sci Res 2012 Feb 2:1373-1384 [FREE Full text]
2. Liu Q, Vasarhelyi M. Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information. 2013 Nov Presented at: 29th world continuous auditing and reporting symposium (29WCARS); November 21-22, 2013; Brisbane, Australia URL: http://raw.rutgers.edu/docs/wcars/29wcars/Health%20care%20fraud%20detection%20A%20survey%20and%20a%20clustering%20model%20incorporating%20Geo-location%20information.pdf
3. Kumar M, Ghani R, Mei ZS. Data mining to predict and prevent errors in health insurance claims processing. USA: Association for Computing Machinery; 2010 Presented at: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2010; Washington, DC, USA p. 65-74 URL: https://dl.acm.org/doi/abs/10.1145/1835804.1835816 [doi: 10.1145/1835804.1835816]
4. Liu J, Bier E, Wilson A, Guerra-Gomez JA, Honda T, Sricharan K, et al. Graph Analysis for Detecting Fraud, Waste, and Abuse in Healthcare Data. AIMag 2016 Jul 04;37(2):33. [doi: 10.1609/aimag.v37i2.2630]
5. Liu S, Yang H, Li J, Kolmanič S. Preliminary Study on the Knowledge Graph Construction of Chinese Ancient History and Culture. Information 2020 Mar 30;11(4):186. [doi: 10.3390/info11040186]
6. Liu Y, Fu ZJ, Li J, Hou L. Generation of medical encyclopedia knowledge graph. Chin J Med Libr Inf Sci Internet 2018 Jun [FREE Full text]
7. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a Health Knowledge Graph from Electronic Medical Records. Sci Rep 2017 Jul 20;7(1):5994. [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
8. Shi L, Li S, Yang X, Qi J, Pan G, Zhou B. Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services. Biomed Res Int 2017;2017:2858423. [doi: 10.1155/2017/2858423] [Medline: 28299322]
9. Yuan K, Deng Y, Chen D, Zhang B, Lei K. Construction techniques and research development of medical knowledge graph. Appl Res Comput 2018 Jul [FREE Full text]

10. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010;17(5):524-527 [FREE Full text] [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
11. Gong F, Chen Y, Wang H, Lu H. On building a diabetes centric knowledge base via mining the web. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):49 [FREE Full text] [doi: 10.1186/s12911-019-0771-6] [Medline: 30961582]
12. Shen L, Sun H, Wang J, Li J. Plotting knowledge graphs for heart failure. Chinese J Med Libr Inf Sci 2019(5):1-5 [FREE Full text] [doi: 10.3969/j.issn.1671?3982.2019]
13. ICD-10 online versions. World Health Organization. URL: https://www.who.int/classifications/icd/icdonlineversions/en/ [accessed 2020-06-20]
14. Zhang Y, Wang X, Hou Z, Li J. Clinical Named Entity Recognition From Chinese Electronic Health Records via Machine Learning Methods. JMIR Med Inform 2018 Dec 17;6(4):e50 [FREE Full text] [doi: 10.2196/medinform.9965] [Medline: 30559093]
15. Jia Y, Xu X. Chinese Named Entity Recognition Based on CNN-BiLSTM-CRF. : IEEE; 2019 Mar 11 Presented at: 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS); November 23-25, 2018; Beijing, China p. 1-4 URL: https://ieeexplore.ieee.org/document/8663820 [doi: 10.1109/icsess.2018.8663820]
16. Yin M, Mou C, Xiong K, Ren J. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. J Biomed Inform 2019 Oct;98:103289. [doi: 10.1016/j.jbi.2019.103289] [Medline: 31541715]
17. Long S, Yuan R, Yi L, Xue L. A Method of Chinese Named Entity Recognition Based on CNN-BILSTM-CRF Model. Singapore: Springer Singapore; 2018 Presented at: ICPCSEE 2018. Communications in Computer and Information Science; September 21-23, 2018; Zhenzhou, China p. 161-175 URL: https://link.springer.com/chapter/10.1007/978-981-13-2206-8_15
18. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. : Association for Computational Linguistics; 2018 Presented at: Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation; 2018; Hong Kong URL: https://www.aclweb.org/anthology/Y18-1061.pdf
19. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. : Association for Computational Linguistics; 2019 Presented at: North American Chapter of the Association for Computational Linguistics; June 3, 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://www.aclweb.org/anthology/N19-1423.pdf
20. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inf Assoc 2010 Jan 01;17(1):19-24. [doi: 10.1197/jamia.m3378]
21. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]
22. Sohn S, Clark C, Halgrim S, Murphy S, Chute C, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 2014;21(5):858-865 [FREE Full text] [doi: 10.1136/amiajnl-2013-002190] [Medline: 24637954]
23. Wang C, Fan J. Medical Relation Extraction with Manifold Models. : Association for Computational Linguistics; 2014 Presented at: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; June 23-25, 2014; Baltimore, MD p. 828-838 URL: https://www.aclweb.org/anthology/P14-1078.pdf [doi: 10.3115/v1/P14-1078]
24. Ben Abacha A, Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach. J Biomed Sem 2011;2(Suppl 5):S4. [doi: 10.1186/2041-1480-2-s5-s4]
25. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
26. Zeng D, Liu K, Chen Y, Zhao J. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. : Association for Computational Linguistics; 2015 Presented at: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; September, 2015; Lisbon, Portugal p. 1753-1762. [doi: 10.18653/v1/D15-1203]
27. Wu Y, Bamman D, Russell S. Adversarial Training for Relation Extraction. 2017 Presented at: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017; Copenhagen, Denmark p. 1778-1783 URL: https://www.aclweb.org/anthology/D17-1187/ [doi: 10.18653/v1/d17-1187]
28. Stoilos G, Stamou G, Kollias S. A String Metric for Ontology Alignment. In: Gil Y, Motta E, Benjamins VR, Musen MA. editors.: Springer; 2005 Presented at: The Semantic Web -- ISWC 2005; November 6-10, 2005; Galway, Ireland p. 624-637 URL: https://link.springer.com/chapter/10.1007/11574620_45#Bib1 [doi: 10.1007/11574620_45]
29. Zhang Y, Paradis T, Hou L, Li J, Zhang J, Zheng H. Cross-Lingual Infobox Alignment in Wikipedia Using Entity-Attribute Factor Graph. : Springer International Publishing; 2017 Presented at: The Semantic Web -- ISWC 2017; October 21-25, 2017; Vienna, Austria p. 745-760. [doi: 10.1007/978-3-319-68288-4_44]
30. Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: a versatile graph matching algorithm and its application to schema matching. 2002 Presented at: Proceedings 18th International Conference on Data Engineering; 2002; San Jose, CA. [doi: 10.1109/icde.2002.994702]
31. Xia Y, Zhao H, Liu K, Zhu H. Normalization of Chinese Informal Medical Terms Based on Multi-field Indexing. In: Zong C, Nie J-Y, Zhao D, Feng Y. editors. Berlin, Heidelberg: Springer; 2014 Presented at: Natural Language Processing and Chinese Computing; December 5-9, 2014; Shen Zhen, China p. 311-320. [doi: 10.1007/978-3-662-45924-9_28]

32. Castaño J, Gambarte M, Park H, Avila WMDP, Pérez D, Campos F, et al. A Machine Learning Approach to Clinical Terms Normalization. Berlin, Germany: Association for Computational Linguistics; 2016 Presented at: Proceedings of the 15th Workshop on Biomedical Natural Language Processing; August, 2016; Berlin, Germany URL: https://www.aclweb.org/anthology/W16-2901/ [doi: 10.18653/v1/w16-2901]
33. Bilenko M. Learnable Similarity Functions and Their Applications to Clustering and Record Linkage. : AAAI Press; 2004 Presented at: Proceedings of the 19th National Conference on Artifical Intelligence; July, 2004; San Jose, CA p. 981-982.
34. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. J Biomed Inform 2015 Dec;58:11-18 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.010] [Medline: 26385377]
35. Angeli G, Tibshirani J, Wu J, Manning C. Combining Distant and Partial Supervision for Relation Extraction. : Association for Computational Linguistics; 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP}); October, 2014; Doha, Qatar p. 1556-1557 URL: https://www.aclweb.org/anthology/D14-1164/ [doi: 10.3115/v1/d14-1164]
36. kg test. URL: https://web.archive.org/web/20191231152615/http://121.12.85.245:1347/kg_test/ [accessed 2019-12-31]

Abbreviations

AT: adversarial training
ATC: Anatomical Therapeutic Classification
BERT: Bidirectional Encoder Representations from Transformations
BiLSTM: bidirectional long short-term memory
BIO: Beginning-Inside-Outside
CNN: convolutional neural network
CRF: conditional random field
DDI: drug-drug interaction
FWA: Fraud, Waste, and Abuse
ICD: International Classification of Diseases
NER: named entity recognition
PCNN: piecewise convolutional neural networks

Edited by B Tang, T Hao, Z Huang; submitted 31.12.19; peer-reviewed by C Friedrich, Z Yang; comments to author 23.02.20; revised version received 13.04.20; accepted 28.05.20; published 23.07.20.

Please cite as:
Sun H, Xiao J, Zhu W, He Y, Zhang S, Xu X, Hou L, Li J, Ni Y, Xie G
Medical Knowledge Graph to Enhance Fraud, Waste, and Abuse Detection on Claim Data: Model Development and Performance Evaluation
JMIR Med Inform 2020;8(7):e17653
URL: http://medinform.jmir.org/2020/7/e17653/
doi:10.2196/17653
PMID:32706714

Original Paper

Temporal Expression Classification and Normalization From Chinese Narrative Clinical Texts: Pattern Learning Approach

Xiaoyi Pan[1], MPhil; Boyu Chen[1], MPhil; Heng Weng[2], PhD; Yongyi Gong[1], PhD; Yingying Qu[3], PhD

[1]School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China

[2]Department of Big Data Research of Medicine, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China

[3]School of Business, Guangdong University of Foreign Studies, Guangzhou, China

Corresponding Author:
Yingying Qu, PhD
School of Business
Guangdong University of Foreign Studies
Faculty Building, Higher Education Mega Center
Guangdong University of Foreign Studies
Guangzhou,
China
Phone: 86 020 3932 8022
Email: jessie.qu@gdufs.edu.cn

Abstract

Background: Temporal information frequently exists in the representation of the disease progress, prescription, medication, surgery progress, or discharge summary in narrative clinical text. The accurate extraction and normalization of temporal expressions can positively boost the analysis and understanding of narrative clinical texts to promote clinical research and practice.

Objective: The goal of the study was to propose a novel approach for extracting and normalizing temporal expressions from Chinese narrative clinical text.

Methods: TNorm, a rule-based and pattern learning-based approach, has been developed for automatic temporal expression extraction and normalization from unstructured Chinese clinical text data. TNorm consists of three stages: extraction, classification, and normalization. It applies a set of heuristic rules and automatically generated patterns for temporal expression identification and extraction of clinical texts. Then, it collects the features of extracted temporal expressions for temporal type prediction and classification by using machine learning algorithms. Finally, the features are combined with the rule-based and a pattern learning-based approach to normalize the extracted temporal expressions.

Results: The evaluation dataset is a set of narrative clinical texts in Chinese containing 1459 discharge summaries of a domestic Grade A Class 3 hospital. The results show that TNorm, combined with temporal expressions extraction and temporal types prediction, achieves a precision of 0.8491, a recall of 0.8328, and a F1 score of 0.8409 in temporal expressions normalization.

Conclusions: This study illustrates an automatic approach, TNorm, that extracts and normalizes temporal expression from Chinese narrative clinical texts. TNorm was evaluated on the basis of discharge summary data, and results demonstrate its effectiveness on temporal expression normalization.

(JMIR Med Inform 2020;8(7):e17652) doi:10.2196/17652

KEYWORDS

Temporal expression extraction; Temporal expression normalization; Machine learning; Heuristic rule; Pattern learning; Clinical text

Introduction

Temporal information, expressions of words or phrases about time, is vital to process and understand data related to time dimension [1]. The automatic extraction of temporal information using natural language processing (NLP) techniques has become a research hotspot [2]. The extraction and summary of events that contain temporal information in chronological order play a key role in many NLP applications such as text summarization [3]. In practice, temporal information can be applied to many tasks (eg, temporal indexing for indicating medical entities of

a united timeline to help comprehend clinical notes and further analysis) [4].

In the medical domain, temporal information has been proved to be useful in clinical research advances and remains essential to the analysis and understanding of clinical events hidden in narrative clinical texts [2]. As typical medical data sources, electronic medical records (EMRs) are collections of electronically stored records that keep medical treatment information of patients in the hospital. These EMRs contain massive unstructured narrative clinical texts (eg, discharge summaries, progress notes). Almost all types of records contain temporal expressions (TEs) as an important indication of clinical information about disease treatment [5,6]. Thus, extraction and normalization of temporal information from these unstructured texts is exceedingly valuable for clinical timeline construction as well as diagnosis procedure identification for clinical analysis [7].

However, the extraction and normalization of clinical temporal information presents difficulties in the current situation. Narrative clinical texts authored by clinical researchers normally includes a large amount of domain terminologies, making them relatively more complicated than other types of narrative texts [8]. Additionally, certain TEs written in different recording habits also increase difficulty of extraction from texts [9]. Most available systems for temporal information processing from EMRs are designed for English texts, while just a small number of them are proposed for Chinese texts [10]. In addition, most shared tasks (eg, the Informatics for Integrating Biology and the Bedside [i2b2] NLP Challenge and several Clinical TempEval Tasks) concentrate on TE extraction and relation identification and seldom pay attention to TE normalization. Relatively speaking, there are several challenges in the task of TE normalization. For instance, TEs are expressed in various formats, causing difficulties in normalization [7]. In addition, certain TEs are dependent on each other, and the normalization needs to identify and compute their reference time. Therefore, time resolution is required to determine whether a TE needs a reference time and how to identify an appropriate reference time for normalization correctly.

Since 2006, multiple clinical NLP shared tasks have been released for open participation, which delivered positive impacts on the development of clinical NLP research [11]. Temporal information processing (including temporal information extraction, temporal relation identification, and TE normalization) was involved in these tasks. In 2012, the sixth i2b2 NLP Challenge concentrated mainly on temporal relation extraction from clinical narratives [12]. A corpus of clinical discharge summaries containing annotated events and TEs was provided in this challenge. The task comprised extraction of (1) clinical events containing medical concepts (eg, clinical departments, tests) and events associated with the clinical timeline of patients including admissions, transfers among different departments, and so on; (2) TEs, including the types of date, time, duration, or frequency (the standardized value of extracted TEs must refer to an International Organization for Standardization [ISO] specification standard); and (3) temporal relations between TEs and clinical events [12]. Other similar tasks were Clinical TempEval 2015 Task 6 [13] and Clinical TempEval 2016 Task 12 [14]. In 2017, SemEval-2017 Task 12 [15] focused on timeline extraction in the clinical domain. This task used pathology reports as well as clinical notes of cancer patients as experiment data and proposed a domain adaptation problem in temporal information processing. Data from colon cancer patients was selected as the training dataset, and data from brain cancer patients was used as the testing dataset. In the task, MacAvaney et al [16] presented a supervised learning approach for TE extraction and event spans, including conditional random fields (CRFs) and decision tree ensembles.

There are clinical EMRs in different languages and a good amount of research on processing temporal information of EMRs. However, most of the research and systems are designed for English TE extraction and normalization. Luo et al [17] developed a method based on CRF for extracting temporal constraints from eligibility criteria in clinical studies. Chang et al [18] applied a method combining regular expression rules, compositional rules, and filtering rules to identify TEs in text. For the purpose of temporal analysis, Tao et al [19] proposed an ontology-based method for temporal information representation of vaccine adverse events. A comprehensive approach consisting of regular expression, pattern matching, and machine learning was developed by Sohn et al [20] for temporal information processing. Kovac˘evic´ et al [21] developed a system in which rules were applied to identification and normalization of TEs and a CRF approach was used for events and temporal identification. In order to identify temporal relationship of entities, Chang et al [22] designed a hybrid method containing a rule-based approach and a maximum entropy model. Sun et al [1] transformed the task of normalizing relative and incomplete temporal expressions (RI-TIMEXes) from narrative clinical texts into a multilabel classification problem and developed a normalization system that included an anchor point classifier, anchor relation classifier, and RI-TIMEX text span parser based on rules. Wang et al [23] presented an approach based on shallow syntactic information and crude properties of extracted event and temporal entities for temporal information tagging and relation extraction. Zhu et al [24] proposed an integrated method based on syntactic parsing for extracting structured medical information and associating temporal information from online health communities. In the method, temporal and medical phrase extraction was regarded as a series of tagging, and temporal relation identification was regarded as a classification problem. Lee et al [25] indicated that the main category of temporal relations is direct temporal relation, which contained significant information required for clinical applications. They constructed a corpus composed of direct temporal relations between events and TEs and proposed an automated support vector machine–based system for direct temporal relation. Meanwhile, research related to Chinese TE extraction and normalization was reported. Wu et al [26] proposed a temporal parser to extract and standardize Chinese TEs. Zhou et al [27] established a framework concentrating on processing narrative clinical records in Chinese, including a regular expression matching–based method for TE identification and an approach for temporal relationship extraction using CRF. Li et al [28] further developed Chinese HeidelTime recourses to solve problems in Chinese temporal tagging (extraction and normalization). For the purpose

of extracting and normalizing TE from Chinese clinical texts, Liu et al [10] designed a system containing a set of rules for each type of TE. Hao et al [9] presented an approach called temporal expression extractor combining heuristic rules with a pattern learning method for TE extraction and normalization in multilingual narrative clinical texts. In general, existing research on TE extraction and temporal relationship extraction has not achieved enough performance for clinical research practice. Particularly, very little research is concentrated on TE normalization from Chinese narrative clinical texts, and most of it is rule-based strategy only.

In this paper, we proposed a hybrid method named TNorm by incorporating a rule-based and a pattern learning-based strategy for TE extraction, classification, and normalization from Chinese narrative clinical texts. TNorm aimed to solve difficulties caused by various formats of TEs and reference time identification for each TE. In TNorm, two groups of patterns were automatically generated from annotated Chinese clinical discharge summaries. The first group was learned and combined with a set of heuristic rules for extracting TEs. After that, TNorm applied a list of extracted temporal features to classify those expressions into a list of temporal types with the help of machine learning algorithms. Finally, combining with rules, the second group of patterns was generated and applied to normalize the identified and classified TEs. The innovation of TNorm was on the combination of rules and pattern learning to solve the difficulties mentioned for Chinese TEs extraction and normalization. In addition, TNorm is compatible with existing classification algorithms to combine the tasks of temporal extraction and normalization. The TEs extracted and normalized by TNorm could be used to generate a corresponding medical events timeline from Chinese narrative clinical texts.

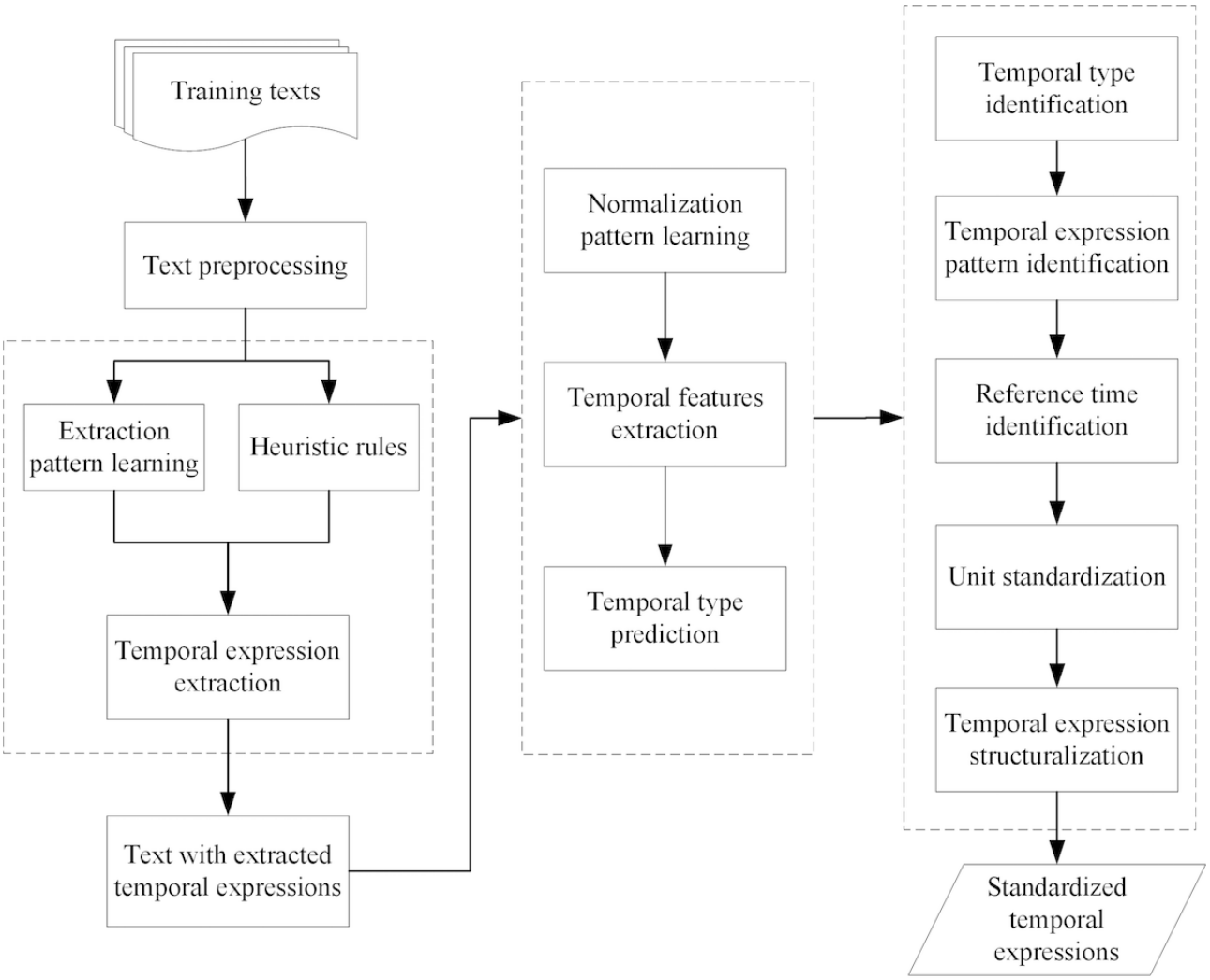
In order to evaluate the performance of the proposed method, we used 1495 unstructured discharge summaries of breast cancer patients from a 3A hospital in China, among which 900 discharge summaries were randomly selected and manually annotated. In temporal type classification, TNorm with a randomizable filtered classifier (RFC) achieved a macro-average F1 score of 0.9573. In the evaluation of normalization, TNorm achieved a precision of 0.8491, recall of 0.8328, and F1 score of 0.8409. The experiment results demonstrated that TNorm has reliable performance on TE extraction and normalization.

Methods

Overall Framework

An automatic method called TNorm was designed for TE extraction and normalization of narrative Chinese clinical texts. It incorporates heuristic rules, automatically learned patterns, and machine learning algorithms and presented a temporal representation as a triple $TE = <M, A, N>$. *TE* denotes a set of temporal mentions as *M*, a set of type attributes of *M* as *A*, and a set of mention values in normalized form as *N*. TNorm transforms TEs into normalized format by referring to two international standards: (1) TimeML [29], a formal specification language for events and TEs, and (2) ISO 8601 [30]. These reference standards are commonly used in many international challenge tasks (eg, 2012 i2b2 Challenge, Clinical TempEval Task). In general, TNorm is proposed to solve the following tasks: (1) extracting temporal mentions *M* from narrative clinical texts, (2) predicting the attributes *A* of mentions *M*, and (3) achieving normalized TE values *N* by standardizing the values of *M*. The framework of TNorm is presented in Figure 1.

Figure 1. Framework of TNorm.



In the extraction process, TEs of available texts are extracted in combination with learned pattern and heuristic rules. Then, in the process of temporal type prediction, temporal features are extracted from the texts with annotated TEs to predict temporal types. Finally, the predicted temporal type and another set of learned patterns are combined in the process of normalization.

To extract TEs, we apply a hybrid approach to deal with narrative Chinese EMRs [9] since this paper mainly focuses on mention type classification and expression normalization. The approach analyzes the annotated TE and summarizes a group of temporal features, by which a list of heuristic rules is established. After that, a list of extraction patterns, used to identify TEs, is automatically learned from clinical training datasets. These extraction patterns are then combined with heuristic rules to extract TEs from clinical texts.

Reference Standard

In the normalization of TEs, two reference standards are adopted: TimeML and ISO 8601. TimeML is applied to define the annotation tags of TEs and ISO 8601 format to standardize the value of TEs.

TimeML includes seven defined tags, <EVENT>, <TIMEX3>, <SIGNAL>, <MAKEINSTANCE>, <TLINK>, <SLINK>, and <ALINK>, which are used to annotate different types of objects. Therefore, we use the tag <TIMEX3> in TNorm to annotate TEs. Accordingly, the attributes defined in <TIMEX3> are divided into nonoptional and optional attributes to represent TEs more accurately. Nonoptional attributes are adopted in TNorm with three indicators, Timex ID number (tid), Type, and Value. In addition, anchorTimeID, an optional attribute for recording the reference TE ID, is used merely when the currently processed TE requires a reference time. For example, a TE is annotated as follows: <TIMEX3 tid="[ID number]" Type="[DATE|DURATION|SET|TIME]" Value="[standardized value of TE]" anchorTimeID="[reference time ID]">[TE]</TIMEX3>.

The value attribute is given referring to ISO 8601 format, which defines a widely accepted representation (eg, "YYYY-MM-DD") for date and time. Based on the representation, ISO 8601 states a series of standardized representation formats. Table 1 shows the formats defined in TNorm.

Table 1. International Organization for Standardization 8601 formats defined in TNorm.

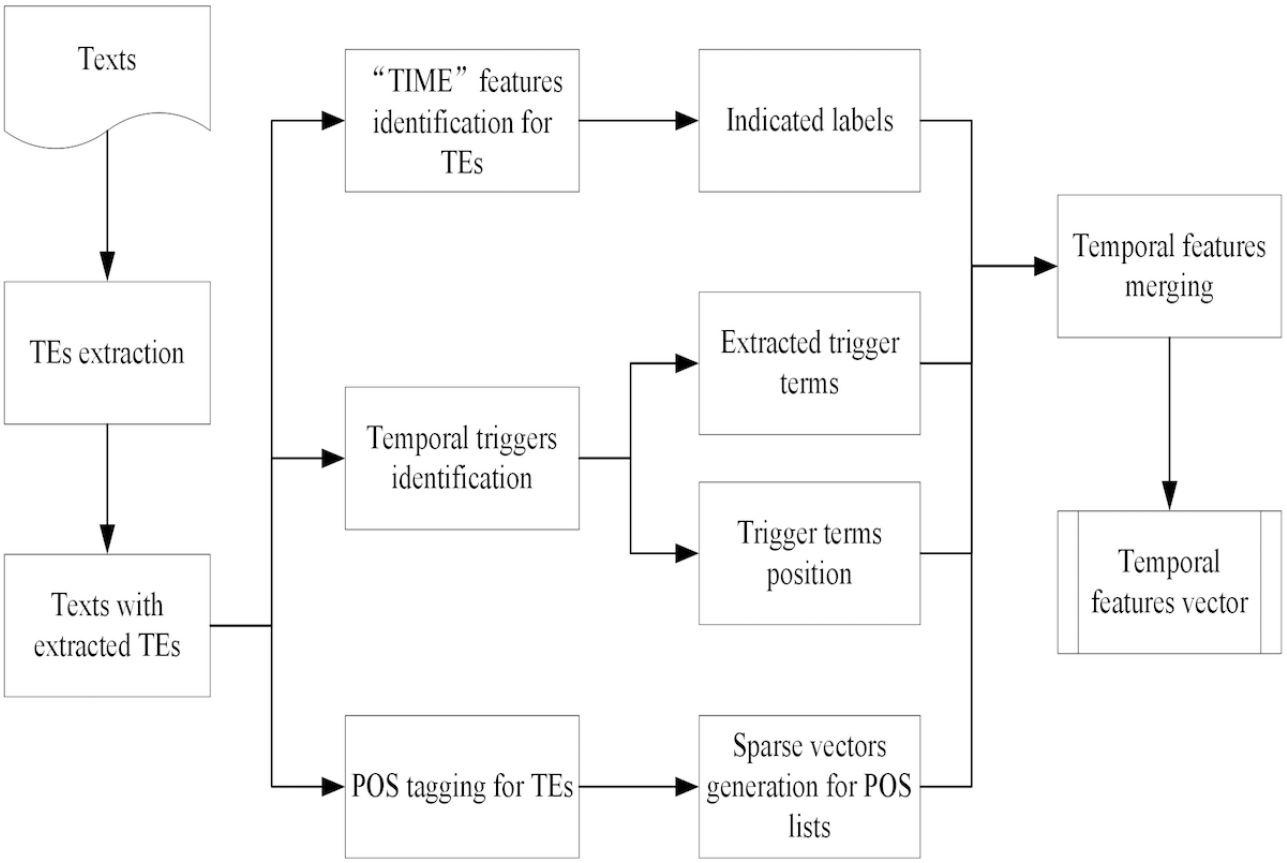
Format	Temporal expression in Chinese	Value
YYYY-MM-DD	2014年6月3日	2014-06-03
YYYY-MM	2014年6月	2014-06
YYYY	2014年	2014
YYYY-MM-DDThh:mm:ss	2014年6月3日上午7点20分4秒	2014-06-03T07:20:04
PnYnMnDTnHnMnS	两年五个月	P2Y5M

Temporal Type Predication

We identified and extracted a group of temporal features to predict the types of extracted TEs by using machine learning algorithms. We treated the temporal type prediction process as a multiclassification task and used TNorm to leverage machine learning algorithms for prediction. In TNorm, the following temporal features are identified and extracted from TEs in clinical training datasets: (1) part-of-speech tags of TEs, processed and generated by Stanford CoreNLP [31]; (2) trigger terms (eg, temporal units); (3) trigger positions, the relative positions of trigger terms; and (4) indicated labels, labels that indicate whether TEs contain typical features of type "TIME." The process of extracting these temporal features is presented in Figure 2.

Through the extraction process of TNorm, TEs in clinical texts are extracted and annotated. In the temporal features extraction process, TNorm identifies typical "TIME" features of the extracted TEs for generating corresponding indicated labels. Through identifying temporal triggers from extracted TEs and their context, TNorm extracts temporal trigger terms and their corresponding positions. In addition, TNorm applied the Stanford CoreNLP to generate corresponding part-of-speech tag lists of the extracted TEs and transform the tags into sparse vectors. Finally, the indicated labels, extracted trigger terms, positions of trigger terms, and sparse vectors are merged into new vectors as the temporal feature vectors of TEs. The feature vectors extracted from the training dataset are processed by machine learning algorithms, which are applied to generate a classification model for temporal type prediction. In this paper, we use the Waikato Environment for Knowledge Analysis (Weka), a machine learning toolkit [32], to use classification algorithms. The initial parameters of classification algorithms set by default in Weka are used in TNorm. After that, classification algorithms predict temporal types of extracted TEs on the testing dataset with temporal feature vectors. We classified the TEs into four types according to TimeML: (1) TIME, (2) DATE, (3) SET, and (4) DURATION. TNorm selects proper normalization process for different TEs based on their temporal types and formats. For instance, regular TEs presented as the DURATION or other temporal types are normalized directly (eg, "一个月", "2014/10/11 7:48:16" and "2014-10-13" are normalized as <TIMEX3 tid="t1" Type="DURATION" Value="P1M">, < TIMEX3 tid="t2" Type="TIME" Value="2014-10-11T07:48:16">, and <TIMEX3 tid="t3" Type="DATE" Value="2014-10-13"> by TNorm).

Figure 2. Flow of temporal features extraction.



Temporal Expression Normalization

To normalize extracted TEs, a list of normalization patterns is automatically learned and generated. For the pattern generation process, a set of candidate patterns is first extracted from the annotated training dataset and then matched back to original texts in the training dataset for validation. The patterns that have confidence scores higher than a predefined threshold are kept. Finally, heuristic rules are summarized through manual observation and applied to normalize the extracted TEs. The learned patterns are used to validate and correct normalized temporal values simultaneously.

After temporal type prediction, TNorm determines reference time for the extracted TEs on the basis of their TE formats and contexts. The reference time plays a key role in transforming the values of TEs into a standard format, so the identification of suitable reference time is important in the normalization process. We proposed three strategies to identify reference time.

When using occurrence time of critical events, certain occurrence times of clinical-related events in narrative texts can be regarded as the reference time if it is highly relevant to current TEs. Through analyzing the context, we classified some events as critical events according to the distance between their locations and TEs in the same sentences. By analyzing the characters in discharge summaries, we found that critical events in an EMR consist of admission, discharge, operation, chemotherapy, and so on. A group of clinical-related events (eg, "回院" [back to the hospital], "化疗后" [after chemotherapy], "化疗后" [postoperation]) are summarized. TNorm detects these events in unstructured clinical texts to acquire corresponding reference time.

When using reference time of special phrases, certain phrases that have a strong relationship with some critical events can be classified as identifiers. We take the time of a critical event as the reference time of all TEs in corresponding special phrases in the same paragraphs. For example, the "入院诊断" (admitting diagnosis) phrase is related with the critical event "入院" (admission), and the reference time of TEs in the paragraph is the occurrence time of "入院" (admission).

When using nearest direct TEs, if the types of TEs are not "DURATION" and they can be normalized directly without reference time, they are defined as direct time that can be applied as the reference time for further normalization. Table 2 shows examples of identifying the reference time of indirect TEs. Figure 3 shows the algorithm for reference time identification.

As shown in Table 2, the critical event of the TE "第*7*、*10*、*14*天" is "化疗" (chemotherapy). The occurrence time of the event can be extracted from narrative texts and thus the reference time of this TE is the date of chemotherapy ("化疗"). For the second example, the TE "*72*小时后" is irrelevant to any critical clinical-related event but is related in a special phrase "出院医嘱" (discharge instruction) whose reference time is the date of discharge. In the EMR, the date of discharge exists, and thus the reference time of this TE is the same as that of the special phrase. Nevertheless, if the reference time of the special phrase is not mentioned in the text, such as in the third example, we choose to identify the nearest direct time.

All TEs are normalized after accomplishing the processes mentioned above. However, using heuristic rules alone may result in incorrect normalization results. Some TEs may connect with more than one medical entity (eg, “主诉:右乳腺癌术后*3*月余,返院行第*7*次化疗” [Chief complaint: More than 3 months after the right breast cancer surgery, patient returned to the hospital for 7th chemotherapy]). The TE is “*3*月余” (more than 3 months) and the medical entities are “右乳腺癌术后” (after the right breast cancer surgery) and “返院” (returned to the hospital). According to the rule-based method, the reference time of this TE is the occurrence time of “返院” (returned to the hospital) and the calculated normalization date is 3 months later than the reference time. However, the correct value of the TE should be equal to the reference time.

To rectify such issues, we applied TNorm to automatically extract a list of patterns from the narrative clinical training dataset with labeled TEs. The detailed procedure of pattern extraction includes the following steps:

1. Label identification: we use the Natural Language Toolkit (NLTK) to split texts into sentences and apply regular expressions to identify TEs that use their reference time as normalization value
2. Temporal label substitution: for retaining contextual information and conveniently extracting patterns, initial TE tags identified in step 1 are substituted by given tags
3. Potential temporal patterns extraction: in the algorithm, given tags and their adjoining words are extracted, and the maximum length of pattern is stipulated. A tag can be contained in several different patterns and all extractive patterns with prescriptive length are regarded as potential patterns
4. Pattern validity verification: the algorithm verifies availability of every potential pattern through applying it to the original dataset and calculates its matching accuracy that can be used as its confidence score
5. Pattern filtration: depending on experiments, the threshold of confidence score is regulated as 0.8. A pattern is adopted merely when its confidence score is higher than the threshold or identical to it. The rest of the patterns are deleted
6. Remove the patterns having substrings: in the filtered pattern group, several patterns are substrings of other patterns. For simplifying the pattern group, a pattern with other substring in the group is removed. Figure 4 shows the algorithm for automatic temporal normalization pattern learning

For instance, when applying the algorithm to an annotated sentence “遂于*<TIMEX3 tid=“t9” Type=“DATE” Value=“2014-09-22” anchorTimeID=“t3”>22/9</TIMEX3>*行右乳癌改良根治术”, the tag and TE are replaced with a given tag (eg, “遂于*<TIMEX3>*行右乳癌改良根治术”). Fifteen patterns are extracted from this sentence, from which the support and confidence scores are calculated. After comparing the threshold and identifying the substring, there are three patterns left, “*<TIMEX3>*办”, “予*<TIMEX3>*”, and “可予*<TIMEX3>*,” which are combined with rules to extract and normalize TEs from texts.

Table 2. Examples of temporal expressions and their corresponding reference time in texts.

Example in text	Temporal expression	Critical event	Special phrase	Reference time
出院医嘱*: 1*、化疗后第 *7*、*10*、*14* 天复查血象*;*	第 *7*、*10*、*14* 天	化疗	出院医嘱	date of chemotherapy
出院医嘱*: 1*、保持伤口清洁干燥,*72* 小时后自行拆除绷带*;*	*72* 小时后	none	出院医嘱	date of discharge
出院情况 : 目前患者第一次化疗结束 , 未诉特殊不适 , 交代相关注意事项后 , 准予出院。	目前	none	出院情况	nearest direct time

Figure 3. Reference time identification.

Algorithm 1 Reference time identification

```
Input a clinical discharge summary text T with temporal expression annotations
temporal formats set format
temporal expressions with annotation tag Items ← Temporal expression identification(T)
n_Text ← T
ForEach item in Items
    type ← temporal type contained in item
    te ← temporal expression contained in item
    reference time ref ← none
    If (reference time requirement (type, format))
        If (critical event identification (te))
            ref ← occurrence time of critical event
        Else If (special phrase identification (te))
            ref ← reference time of special phrase
        Else
            ref ← nearest direct temporal expression
    EndIf
    value ← temporal normalization (te, ref)
    n_item ← new annotation generation (item, value)
    n_Text ← replacement (n_Text, item, n_item)
EndFor
Output n_Text
```

Figure 4. Automatic temporal normalization pattern learning.

Algorithm 2 Automatic Temporal Normalization Pattern Learning

Input a clinical discharge summary text *T* with temporal expressions normalized
temporal expressions with annotation tag *Items* ← Temporal expression identification(*T*)
final pattern set *f_patts* ← 0; pattern set *patts* ← 0; potential pattern set *p_patts* ← 0
ForEach *item* in *Items*
  If (anchorTimeID identification (item))
    Value, tid ← temporal normalization value, temporal id value contained in *item*
    anchorID ← anchorTimeID value contained in *item*
    anchorValue ← anchor temporal expression value identification (*T, anchorID*)
    If (*value* equals to *anchorValue*)
      sen ← sentence identification and tag substitution (*T, tid,* original tag, specified tag)
      word window size *wl* ← 8
      ForEach (*w_window* in windows from *sen* with size *wl*)
        generate structures as potential patterns from *w_window*
        add the patterns to *p_patts* if they are >= two words
      EndFor
    EndIf
  EndIf
  ForEach (*p_patt* in *p_patts*)
    c_match, t_match ← the number of correct matches, total match in match (*p_patt*)
    If (# *c_match* >= support threshold *st*)
      confidence ← *c_match* / *t_match*
      If (*confidence* >= confidence threshold *ct*)
        add (*p_patts*, *confidence*) to *patts*
  ForEach (*patt* in *patts*)
    (*sub_patt*, *sub_confidence*) ← substring identification (*patt*, *patts*)
    If (*sub_patt* is None or *sub_confidence* is not equal to the confidence of *patt*)
      add *patt* to *f_patts*
    EndIf
  EndFor
EndFor
Output *f_patts*

Results

Datasets

The experiment dataset contains 1459 Chinese discharge summary texts of patients with breast cancer from a 3A hospital in mainland China. We randomly selected 900 EMRs and manually annotated all TEs. A TE was labeled with tag "<TIMEX3></TIMEX3>" and normalized as the standard form for evaluation. According to the TimeML standard, a label needs to contain four attributes: (1) tid, the index of temporal information in the record; (2) Type, the temporal type; (3) Value, the normalized date of a TE; and (4) anchorTimeID, the index of the reference time of the current TE. Each TE has a unique tag (eg, <TIMEX3 tid="t1" Type="TIME" Value="2014-10-11T07:48:16">2014/10/11 7:48:16</TIMEX3>, <TIMEX3 tid="t7" Type="DATE" Value="2014-08-02" anchorTimeID="t2">1 周前</TIMEX3>, <TIMEX3 tid="t4" Type="DURATION" Value="P1Y"> 1年余</TIMEX3>, <TIMEX3 tid="t13" Type="SET" Value="[10-29]" anchorTimeID="t11"> 第7、9、14 天 </TIMEX3>). These 900 EMRs contain 12,096 TEs (13.44 TEs per text). Details of the training and testing datasets are illustrated in Table 3.

Table 3. Statistics of the dataset containing Chinese discharge summary texts.

Datasets	Texts, n	Temporal expressions, n	Temporal expressions per text, mean
Training	450	5966	13.26
Testing	450	6130	13.62
Total	900	12,096	13.44

Evaluation Metrics

The task of temporal type prediction can be regarded as a multiclassification task. Since a classification algorithm may display different classification capabilities on different types of TEs and the number of temporal types is rather different in a text, the traditional precision, recall, and F1 are not appropriate to indicate the actual classification capability of a classification algorithm. For instance, a classification algorithm may obtain high precision of "TIME" (one temporal type) prediction but poor precision of "SET" (one temporal type) prediction. It will perform better in a dataset that contains more "TIME" than "SET" and worse in the opposite condition. Therefore, to reduce the differentiation, we applied macro-average precision, macro-average recall, and macro-average F1-measure rather than the traditional precision, recall, and F1 to evaluate the prediction result, as shown in the following equations (Figure 5).

In the equations, n represents the quantity of labels while k is a number that represents different labels. P_k and R_k stand for the precision and recall, respectively, of the label that corresponds to k.

A normalized TE is considered correct only when the extracted expression and its normalized value are completely identical to the manually annotated result. In the comparison of normalization performance, the metrics precision, recall, and F1-measure are used to evaluate the performance of TNorm. Based on the definition, calculations of the metrics are shown in the following equations (Figure 6).

$\#C_{value}$ represents the number of TEs that are correctly extracted and normalized (with correct value of the attribute "Value"), $\#N_{TE1}$ represents the number of TEs that are normalized by TNorm, and $\#N_{TE2}$ represents the number of TEs that should be normalized in the dataset.

Figure 5. Calculation equations of evaluation metrics macro-average precision, macro-average recall, and macro-average F1-measure. MP: macro-average precision; MR: macro-average recall; MF: macro-average F1-measure.

$$MP = \Sigma_{k=1}^{n} P_k / n \quad (1)$$
$$MR = \Sigma_{k=1}^{n} R_k / n \quad (2)$$
$$MF = (MP + MR)/2 \quad (3)$$

Figure 6. Calculation equations of evaluation metrics precision, recall and F1-measure.

$$Precision = \#C_{value} / \#N_{TE1} \quad (4)$$
$$Recall = \#C_{value} / \#N_{TE2} \quad (5)$$
$$F1 = 2 * Precision * Recall / (Precision + Recall) \quad (6)$$

Outcome

For temporal type prediction, TNorm extracted temporal features vectors (as mentioned in the Methods section) from the training dataset and combined machine learning algorithms with these feature vectors. Since the process of temporal type prediction could be treated as the multiclassification tasks, we used Weka to classify temporal types with 10-fold cross-validation. In a classification task, the main influence factors were extracted features and selected classification algorithms. Thus, the performance of a list of algorithms including logistic, decision table, and k-nearest neighbor (KNN) was calculated and ranked by macro-averaged F1-measure (top 10 only). As the result reported in Table 4 shows, RFC achieved the best F1-measure (0.9573). RFC is a variant of a simple filtered classifier that requires either the base learner or the filter to implement a random interface. In this experiment, the base classifier selected in RFC was KNN(k=1) and the filter selected in RFC was random projection, which was used to reduce the dimensionality of vectors. As shown in the table, the KNN algorithm could perform well individually. Through analysis, we determined that the types, relative position, and context of TEs in each discharge summary were mostly similar. Therefore, most initial generated temporal feature vectors with the same temporal types were close in distance. However, the vectors processed by the filter were probably closer in distance with reducing invalid attributes. As a consequence, RFC that combined KNN with random projection in this experiment achieved high performance and was selected as the baseline machine learning algorithm in TNorm, used in the following normalization procedure.

The main parts of the normalization process with TNorm included heuristic rules and pattern learning. For verifying the validity of pattern learning in temporal normalization, the temporal type prediction process and temporal normalization process (as mentioned in Methods) were applied without the application of automatic temporal extraction. We used the same classification algorithm to predict temporal types and compared the effectiveness of the method under two conditions: with rules only and with both rules and the generated patterns. In the TE normalization task, the normalization result of each TE was

unique. The normalization result generated by the approach could only be divided into right and wrong. Therefore, the evaluation metric Accuracy was used.

In the metric Accuracy = #Correct / $\#N_{TE}$, #Correct represented the number of TEs with correct value of the attributes "Type" and "Value" in the testing dataset and $\#N_{TE}$ represented the number of TEs that should be normalized in the testing dataset. The top 5 classification algorithms were respectively combined with the two method models: rule only and rule plus pattern, which were contained in TNorm to normalize TEs. Based on the testing dataset containing 6130 TEs, as the result shows in Table 5, the strategy of method with rules achieved an accuracy of 0.8587, while the second condition with both the rules and pattern learning achieved an accuracy of 0.8654, demonstrating a positive influence of the learned patterns in the normalization process.

Patterns and temporal features were generated from the training dataset, which indicated that the scale of training dataset might influence the effectiveness of extraction and normalization. We used different sizes of the training dataset to discover how the scale of training dataset affected the performance. In the experiment, the number of EMRs in the testing dataset remained the same (450), while the number of training datasets increased from 50 to 450. The experiment result is presented by a line chart in Figure 7. The experiment result showed that the scale of training dataset had slight influence on the performance of the approach. When the number of EMRs from the training dataset reached 300, the effectiveness of the training dataset tended to be stable. Through analyzing the dataset and experiment result, we found the types, relative positions, and contexts of TEs in each discharge summary were mostly similar. With the number of EMRs in training dataset increasing, the number of learned patterns and generated temporal feature vectors increased. However, the influence of the same patterns and vectors on TNorm was steady and independent of their quantity. In addition, we also tested the stability of TNorm using different sizes of testing datasets (50 to 450) but kept the training dataset the same (450). The result, as Figure 8 shows, illustrates that TNorm reached a comparative stable performance when the number of EMRs was larger than 350. In this experiment, when the testing dataset was increased to 450 records, TNorm achieved a precision of 0.8491, a recall of 0.8328, and an F1 score of 0.8409. With the testing dataset scale increasing, the performance of the F1 score changed slightly.

Table 4. Detailed experiment result of the top 10 classification algorithms.

Classification algorithm	Macro-average precision	Macro-average recall	Macro-average F1
Multiclass classifier	0.9553	0.9420	0.9485
Logistic	0.9558	0.9425	0.9488
Simple logistic	0.9560	0.9423	0.9490
Iterative classifier optimizer	0.9493	0.9525	0.9510
Logit boost	0.9493	0.9525	0.9510
Decision table	0.9493	0.9538	0.9513
JRip	0.9523	0.9518	0.9523
K-nearest neighbor (k=1)	0.9518	0.9613	0.9563
Logistic model trees	0.9545	0.9598	0.9570
Randomizable filtered classifier	0.9535	0.9613	0.9573

Table 5. Evaluation result of the efficiency of the learned patterns in TNorm.

Strategy	Accuracy
Randomizable filtered classifier	
rule	0.8587
rule+pattern	0.8654
Logistic model trees	
rule	0.8587
rule+pattern	0.8654
K-nearest neighbor (k=1)	
rule	0.8587
rule+pattern	0.8654
JRip	
rule	0.8586
rule+pattern	0.8653
Decision table	
rule	0.8586
rule+pattern	0.8653

Figure 7. Performance changes using the method with different sizes of training dataset.

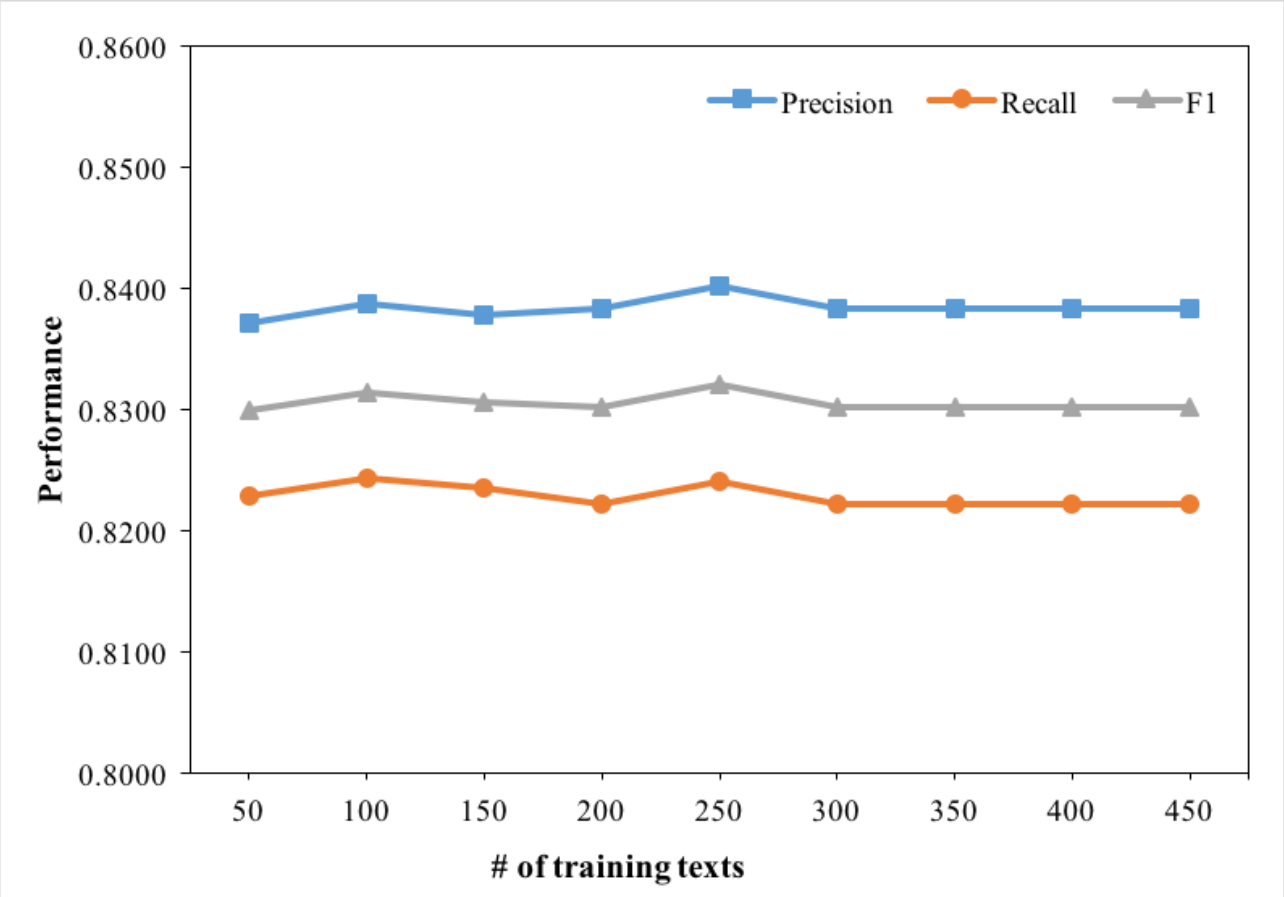
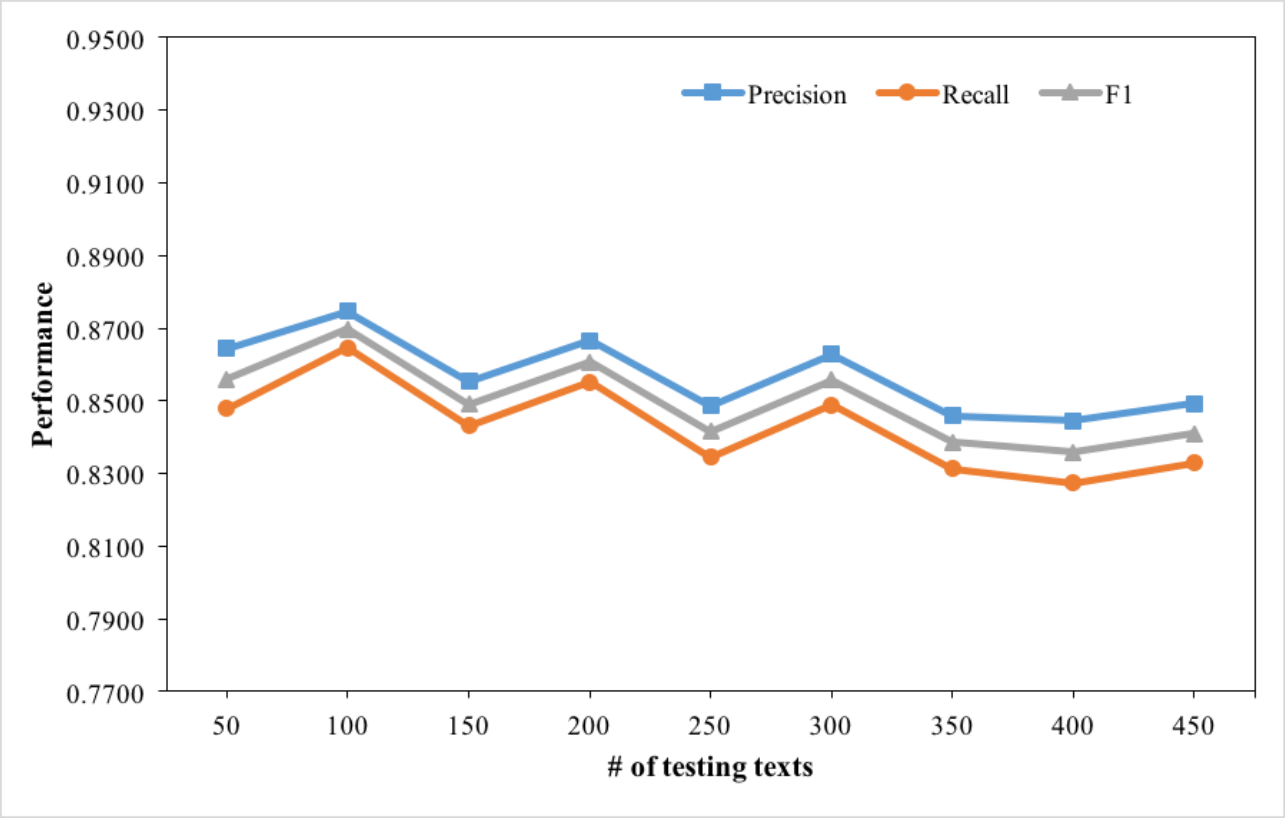


Figure 8. Experiment result of the stability test.



Discussion

Principal Findings

In the experiment of comparing classification algorithms for temporal type prediction, we used the macro-average F1-measure as the ultimate evaluation metric and finally selected RFC as the best to be integrated in TNorm. Throughout the experiment, TNorm performed well on the combination task of TE extraction, temporal type prediction, and temporal normalization. However, a few errors still occurred in each process of applying TNorm.

All error cases from TNorm were analyzed and classified to three types. The first was errors caused by wrong representations or typos in original narrative texts. For instance, “出院日期” (discharge date) followed by the temporal information of discharge date was normally mentioned in every clinical discharge summary. However, in some special discharge summaries, the date was incorrectly written as “出院日期:出院日期” (discharge date: discharge date), which caused the problem of lacking specific discharge date information. As a result, in this wrongly representative text, the TE that required the discharge date as reference time could not be normalized correctly. In addition, some clinical texts contained a series of TEs (eg, the text “门诊 星期一 星期二 星期三 星期四 星期五 上午” [Outpatient Monday Tuesday Wednesday Thursday Friday AM]) without represented any specific time, causing difficulty in normalization.

The second error type was caused by machine learning algorithms for temporal type prediction. Since the rules of temporal normalization were associated with temporal types, TEs with wrong temporal types might be matched with inappropriate rules, causing negative effects in the normalization. For example, in the text “肿物增大*2*月” example, the text “*2*月” was classified as the type “DURATION,” but the classification algorithm predicted it as “DATE.” The incorrect type label resulted in false temporal value (eg, correct normalization result should be “肿物增大<TIMEX3 tid=“t7” Type=“DURATION” Value=“P2M”> *2*月</TIMEX3>” while the result generated by TNorm was “肿物增大<TIMEX3 tid=“t7” Type=“DATE” Value=“2014-2” anchorTimeID=“t3”> *2*月</TIMEX3>”).

The third error type was caused by automatically generated patterns. Although the learned patterns could improve the precision of TNorm, they might cause matching mistakes in special cases. For example, the pattern “ 于我” matched matched the text “*6*天前于我院行双乳*B*超,” thus a temporal value “2014-10-23,” which was the same as the normalized value of its reference time, was computed. However, the correct standardized value of the TE “*6*天前” was “2014-10-17.”

Limitations

There was a limitation of the proposed method. The TNorm consisted of sequential functions including (1) TE extraction, (2) temporal type prediction, and (3) TE normalization. Since time expressions were processed step by step in a sequence order, any errors generated from a step in the process might have negative effects in the next step. To reduce or eliminate this kind of effect, we will try to explore a joint model that conducts the three tasks of extraction, predication, and normalization simultaneously in the future.

Conclusions

This paper proposed a method, TNorm, for automatically extracting and normalizing TEs from Chinese narrative clinical texts. TNorm was composed of alternative machine learning methods, a pattern learning method, and a set of heuristic rules. Several experiments based on 1459 Chinese clinical texts from a 3A hospital in mainland China were conducted to evaluate the performance of classification algorithms, effectiveness of pattern learning, and stability of TNorm, respectively. Results demonstrated that TNorm was reliable and stable for TE normalization of Chinese EMR records.

Acknowledgments

The work was supported by grants from the Scientific and Technology Plan of Guangzhou Project (No. 201804010296, 201904010228, and 201803010063), National Natural Science Foundation of China (No. 61871141), Guangzhou Science Technology and Innovation Commission (No. 201803010063), and Natural Science Foundation of Guangdong Province (No. 2018A030310051).

Conflicts of Interest

None declared.

References

1. Sun W, Rumshisky A, Uzuner O. Normalization of relative and incomplete temporal expressions in clinical narratives. J Am Med Inform Assoc 2015 Oct;22(5):1001-1008. [doi: 10.1093/jamia/ocu004] [Medline: 25868462]
2. Lee H, Xu H, Zhang Y, Moon S. UTHealth at SemEval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. 2016 Presented at: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016); 2016; San Diego. [doi: 10.18653/v1/S16-1201]
3. Strötgen J, Gertz M. Heideltime: high quality rule-based extraction and normalization of temporal expressions. 2010 Presented at: Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics; 2010; Los Angeles p. 321-324 URL: https://www.aclweb.org/anthology/S10-1071.pdf
4. Liu Z, Wang X, Chen Q, Tang B, Xu H. Temporal indexing of medical entity in Chinese clinical notes. BMC Med Inform Decis Mak 2019 Jan 31;19(Suppl 1):17 [FREE Full text] [doi: 10.1186/s12911-019-0735-x] [Medline: 30700331]
5. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc 2011;18(5):594-600 [FREE Full text] [doi: 10.1136/amiajnl-2011-000153] [Medline: 21846787]
6. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017 Dec;73:14-29 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
7. Madkour M, Benhaddou D, Tao C. Temporal data representation, normalization, extraction, and reasoning: a review from clinical domain. Comput Methods Programs Biomed 2016 May;128:52-68 [FREE Full text] [doi: 10.1016/j.cmpb.2016.02.007] [Medline: 27040831]
8. Hao T, Rusanov A, Weng C. Extracting and normalizing temporal expressions in clinical data requests from researchers. 2013 Presented at: International Conference on Smart Health; 2013; Berlin p. 41-51. [doi: 10.1007/978-3-642-39844-5_7]
9. Hao T, Pan X, Gu Z, Qu Y, Weng H. A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts. BMC Med Inform Decis Mak 2018 Mar 22;18(Suppl 1):22 [FREE Full text] [doi: 10.1186/s12911-018-0595-9] [Medline: 29589563]
10. Liu Z, Tang B, Wang X, Chen Q, Li H, Bu J, et al. CMedTEX: a rule-based temporal expression extraction and normalization system for Chinese clinical notes. AMIA Annu Symp Proc 2016:818-826 [FREE Full text] [Medline: 28269878]
11. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
12. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013;20(5):806-813 [FREE Full text] [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]
13. Bethard S, Derczynski L, Savova G. SemEval-2015 Task 6: Clinical TempEval. 2015 Presented at: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); 2015; Denver p. 806-814. [doi: 10.18653/v1/S15-2136]
14. Bethard S, Savova G, Chen WT. SemEval-2016 Task 12: Clinical TempEval. 2016 Presented at: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016); 2016; San Diego p. 1052-1062. [doi: 10.18653/v1/s16-1165]
15. Bethard S, Savova G, Palmer M. SemEval-2017 Task 12: Clinical TempEval. 2017 Presented at: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017); 2017; Vancouver p. 565-572. [doi: 10.18653/v1/s17-2093]
16. MacAvaney S, Cohen A, Goharian N. GUIR at SemEval-2017 Task 12: A Framework for Cross-Domain Clinical Temporal Information Extraction. 2017 Presented at: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017); 2017; Vancouver p. 1024-1029. [doi: 10.18653/v1/s17-2180]

17. Luo Z, Johnson SB, Lai AM, Weng C. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. 2011 Presented at: AMIA Annu Symp; October 22-26, 2011; Washington p. 843-852 URL: http://europepmc.org/abstract/MED/22195142
18. Chang AX, Manning CD. Sutime: a library for recognizing and normalizing time expressions. 2012 Presented at: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12); 2012; Istanbul p. 3735-3740.
19. Tao C, He Y, Yang H, Poland GA, Chute CG. Ontology-based time information representation of vaccine adverse events in VAERS for temporal analysis. J Biomed Semantics 2012 Dec 20;3(1):13 [FREE Full text] [doi: 10.1186/2041-1480-3-13] [Medline: 23256916]
20. Sohn S, Wagholikar KB, Li D, Jonnalagadda SR, Tao C, Komandur Elayavilli R, et al. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. J Am Med Inform Assoc 2013;20(5):836-842 [FREE Full text] [doi: 10.1136/amiajnl-2013-001622] [Medline: 23558168]
21. Kovacevic A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. J Am Med Inform Assoc 2013;20(5):859-866 [FREE Full text] [doi: 10.1136/amiajnl-2013-001625] [Medline: 23605114]
22. Chang Y, Dai H, Wu JC, Chen J, Tsai RT, Hsu W. TEMPTING system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. J Biomed Inform 2013 Dec;46 Suppl:S54-S62 [FREE Full text] [doi: 10.1016/j.jbi.2013.09.007] [Medline: 24060600]
23. Wang W, Kreimeyer K, Woo EJ, Ball R, Foster M, Pandey A, et al. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. J Biomed Inform 2016 Aug;62:78-89 [FREE Full text] [doi: 10.1016/j.jbi.2016.06.006] [Medline: 27327528]
24. Zhu L, Yang H, Yan Z. Extracting temporal information from online health communities. 2017 Presented at: Proceedings of the 2nd International Conference on Crowd Science and Engineering. ACM; 2017; Beijing. [doi: 10.1145/3126973.3126975]
25. Lee H, Zhang Y, Jiang M, Xu J, Tao C, Xu H. Identifying direct temporal relations between time and events from clinical notes. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):49 [FREE Full text] [doi: 10.1186/s12911-018-0627-5] [Medline: 30066643]
26. Wu M, Li W, Lu Q. CTEMP: A Chinese temporal parser for extracting and normalizing temporal information. 2005 Presented at: International Conference on Natural Language Processing; 2005; Berlin p. 694-706. [doi: 10.1007/11562214_61]
27. Zhou XJ, Li HM, Lu XD. Temporal expression recognition and temporal relationship extraction from Chinese narrative medical records. 2011 Presented at: 5th International Conference on Bioinformatics and Biomedical Engineering. IEEE; 2011; Wuhan p. 1-4. [doi: 10.1109/icbbe.2011.5780699]
28. Li H, Strötgen J, Zell J. Chinese temporal tagging with HeidelTime. 2014 Presented at: 14th Conference of the European Chapter of the Association for Computational Linguistics; Apr 26-30; Gothenburg p. 133-137 URL: https://www.aclweb.org/anthology/E14-4026.pdf [doi: 10.3115/v1/e14-4026]
29. Pustejovsky J. TimeML: Robust specification of event and temporal expressions in text. In: New Directions in Question Answering. Netherlands: Kluwer Academic Publishers; 2003:28-34.
30. International Standardization Organization. ISO 8601: Data elements and interchange formats. Information interchange. Representation of dates and times. International Standardization Organization 2004. [doi: 10.3403/03234467]
31. Manning C. The Stanford CoreNLP natural language processing toolkit. 2014 Presented at: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations; 2014; Baltimore p. 55-60. [doi: 10.3115/v1/p14-5010]
32. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004 Apr 08;20(15):2479-2481. [doi: 10.1093/bioinformatics/bth261]

Abbreviations

CRF: conditional random field
EMR: electronic medical record
ISO: International Organization for Standardization
KNN: k-nearest neighbor
NLP: natural language processing
NLTK: Natural Language Toolkit
RFC: randomizable filtered classifier
RI-TIMEX: relative and incomplete temporal expression
TE: temporal expression
tid: Timex ID number
Weka: Waikato Environment for Knowledge Analysis

Edited by T Hao, B Tang, Z Huang; submitted 31.12.19; peer-reviewed by T Hao, Z Li, F Shen; comments to author 14.02.20; revised version received 28.02.20; accepted 13.03.20; published 27.07.20.

Please cite as:
Pan X, Chen B, Weng H, Gong Y, Qu Y
Temporal Expression Classification and Normalization From Chinese Narrative Clinical Texts: Pattern Learning Approach
JMIR Med Inform 2020;8(7):e17652
URL: https://medinform.jmir.org/2020/7/e17652
doi:10.2196/17652
PMID:32716307

Original Paper

Document-Level Biomedical Relation Extraction Using Graph Convolutional Network and Multihead Attention: Algorithm Development and Validation

Jian Wang[1], PhD; Xiaoyu Chen[1], BSc; Yu Zhang[1], MSc; Yijia Zhang[1], PhD; Jiabin Wen[2], PhD; Hongfei Lin[1], PhD; Zhihao Yang[1], PhD; Xin Wang[1], BSc

[1]School of Computer Science and Technology, Dalian University of Technology, Dalian, China

[2]Department of VIP, The Second Hospital of Dalian Medical University, Dalian, China

Corresponding Author:
Yijia Zhang, PhD
School of Computer Science and Technology
Dalian University of Technology
No. 2 Linggong Road, Ganjingzi District
Dalian, 116023
China
Phone: 86 0411 84708498
Email: zhyj@dlut.edu.cn

Abstract

Background: Automatically extracting relations between chemicals and diseases plays an important role in biomedical text mining. Chemical-disease relation (CDR) extraction aims at extracting complex semantic relationships between entities in documents, which contain intrasentence and intersentence relations. Most previous methods did not consider dependency syntactic information across the sentences, which are very valuable for the relations extraction task, in particular, for extracting the intersentence relations accurately.

Objective: In this paper, we propose a novel end-to-end neural network based on the graph convolutional network (GCN) and multihead attention, which makes use of the dependency syntactic information across the sentences to improve CDR extraction task.

Methods: To improve the performance of intersentence relation extraction, we constructed a document-level dependency graph to capture the dependency syntactic information across sentences. GCN is applied to capture the feature representation of the document-level dependency graph. The multihead attention mechanism is employed to learn the relatively important context features from different semantic subspaces. To enhance the input representation, the deep context representation is used in our model instead of traditional word embedding.

Results: We evaluate our method on CDR corpus. The experimental results show that our method achieves an F-measure of 63.5%, which is superior to other state-of-the-art methods. In the intrasentence level, our method achieves a precision, recall, and F-measure of 59.1%, 81.5%, and 68.5%, respectively. In the intersentence level, our method achieves a precision, recall, and F-measure of 47.8%, 52.2%, and 49.9%, respectively.

Conclusions: The GCN model can effectively exploit the across sentence dependency information to improve the performance of intersentence CDR extraction. Both the deep context representation and multihead attention are helpful in the CDR extraction task.

(JMIR Med Inform 2020;8(7):e17638) doi:10.2196/17638

KEYWORDS

biomedical relation extraction; dependency graph; multihead attention; graph convolutional network

Introduction

Valuable biomedical information and knowledge are still hidden in the exponentially increasing biomedical literature, such as the chemical-disease relation (CDR). Extracting the relation between chemicals and diseases is an important task in biomedical text mining, which plays an important role in various biomedical research studies, such as clinical treatment, drug development, and biomedical knowledge discovery [1-3]. However, extracting CDR from the biomedical literature manually is time-consuming and difficult to keep up-to-date. Thus, the BioCreative V community [4] proposed a task of extracting CDR in the biomedical literature automatically to promote the research on the CDR extraction.

To date, many methods have been proposed for automatic relation extraction between chemicals and diseases, which can be divided into 3 categories: rule-based methods [5], feature-based methods [6-9], and deep neural network-based methods [10-13]. Rule-based methods aim to formulate the heuristic rules for CDR extraction. Lowe et al [5] developed a pattern-based system with some heuristic rules to extract chemical-induced disease (CID) relations within the same sentence. The heuristic rules are used to extract the most likely CID relations when no patterns match a document. Generally, rule-based methods are simple and effective. However, these methods are difficult for application in a new task or dataset. Feature-based methods aim at designing rich features, including semantic and syntactic information. Xu et al [6] utilized text features, including context information and entity information, incorporated with domain knowledge to extract CID relations. Since the syntactic information carried in the dependency graph of the sentence is crucial to CDR extraction, some studies also developed syntactic features. Gu et al [7] utilized various linguistic features to extract CID relations with the maximum entropy model. They leveraged lexical features for both intrasentence and intersentence level relation extraction and developed the dependency features only for intrasentence level relation extraction. Zhou et al [8] utilized the shortest dependency path between chemical and disease entities to extract structured syntactic features. Feature-based methods achieve better performance than rule-based methods. However, traditional feature-based methods only use the dependency trees to extract local syntactic dependencies for the intrasentence level relation extraction, without considering the syntactic dependencies across sentences for the document-level relation extraction. Besides, designing rich features is a time-consuming and laborious task.

In recent years, the deep neural network has been widely used in various natural language processing (NLP) tasks. Some studies have developed deep neural network-based methods for biomedical relation extraction. Long short-term memory (LSTM) models and convolutional neural network (CNN) models are the 2 major neural networks. Zhou et al [10] applied LSTM and CNN models based on traditional word embedding to capture context features for CDR extraction and achieve a good performance. Gu et al [11] proposed a CNN-based model to capture context and dependency features for intrasentence level relation extraction. Nguyen and Verspoor [13] investigated character-based word embedding into the CNN-based relation extraction model. Traditional word embedding such as word2vec cannot vary according to linguistic contexts effectively. Peters et al [14] proposed deep contextualized word representations called ELMo based on a deep bidirectional language model. ELMo can generate a more comprehensive representation for each word based on the sentence context. Therefore, integrating ELMo with a deep neural network may improve the performance of CDR extraction.

In both CNN-based and LSTM-based models, it is hard to distinguish the relevant and irrelevant context features for the relation extraction. A recent study [15] suggested that attention mechanism can capture the most important semantic information for the relation extraction. Vaswani et al [16] introduced a multihead attention mechanism that applied the self-attention mechanism multiple times to capture the relatively important features from different representation subspaces. Thus, multihead attention mechanism can be used to improve the performance of the CDR extraction.

Dependency trees are often used to extract local dependencies for intrasentence level CDR extraction. However, existing studies ignored the nonlocal dependency across sentences, which is crucial for intersentence level CDR extraction. Quirk et al [17] introduced a document graph that can derive features within and across sentences. Thus, we also constructed a document-level dependency graph that can extract dependencies for intrasentence and intersentence level CDR extraction simultaneously. Recently, the graph convolution network (GCN) [18] has been effectively used for encoding document graph information. Thus, GCN can operate directly on the document-level dependency graph to capture long-range syntactic information, which is useful for CDR extraction.

In this study, we evaluated the effectiveness of the deep contextualized word representations, multihead attention mechanism, and GCN in the CDR extraction task. To improve the performance of the intersentence relation extraction, we constructed the document-level dependency graph to capture the dependency syntactic information across sentences. Based on the document-level dependency graph, we proposed a novel end-to-end model to extract CID relations from the biomedical literature. First, we used ELMo, POS embedding, and position embedding to construct the input representation and employed the multihead attention with bidirectional LSTM (BiLSTM) to capture the relatively important context features. Second, we employed the GCN to capture the long-range dependency features based on the document-level dependency graph. Third, we combined the context features and long-range dependency features as the final feature representation and applied a *Softmax* function to implement relation classification. Finally, we evaluated our model on the CDR corpus.

Methods

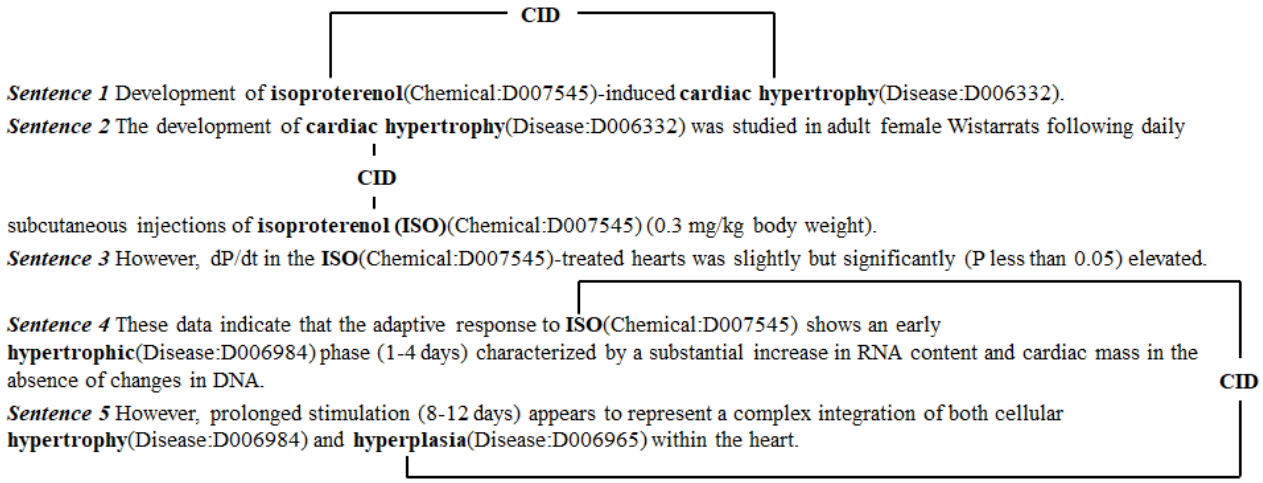
CDR Extraction

The CDR extraction task is a challenging task, which was proposed by the BioCreative V community. The CDR extraction task aims to extract CDR from the biomedical literature

automatically and accurately. It is composed of 2 subtasks: (1) disease named entity recognition and normalization and (2) CID relation extraction.

In this study, we focused on the CID relation extraction task. The CDR extraction task is a document-level biomedical relation extraction problem, which is different from traditional biomedical relation extraction task. Traditional biomedical relation extraction only considers relation within a single sentence such as protein-protein interaction [19] and drug-drug interaction [20]. However, the CID relation is not only expressed within a single sentence, but it is also expressed across several sentences. Figure 1 shows an illustration of CDR extraction. It is extracted from the CDR corpus whose PMID is 6203632. Among these sentences, the texts in bold mention the chemical and disease entities. In Figure 1, we mark the corresponding entity type and the medical subject headings concept identifiers [21] after the entity mention in the sentence. The chemical D007545 has 2 intrasentence level co-occurrences with disease D006332 in the *sentence 1* and the *sentence 2*, while it has an intersentence level co-occurrence with disease D006965. However, not all occurrences of the chemicals and diseases are considered as a CID relation. For example, the chemical D007545 does not have a CID relation with the disease D006984 in the *sentence 4* because the concept of the disease D006984 is too general to reflect a CID relation.

Figure 1. Illustrative examples of CID relation. CID: chemical-induced disease.



Relation Instance Construction

First, we should construct relation instances for both training and testing stages. All the instances generated from the disease and chemical mentions in the document are pooled into 2 groups at the intrasentence and intersentence levels, respectively. The former means that a chemical-disease mention pair is in the same sentence. The latter means that a mention pair is in a different sentence. If the relation between the chemical and disease entity of the mentioned pair is annotated as a CID relation in the document, then this mentioned pair is constructed as a positive instance; otherwise, this mentioned pair is constructed as a negative instance. We applied several effective heuristic rules for both intrasentence and intersentence level instances. The details are as follows.

Relation Instance Construction for Intrasentence Level

1. All chemical-disease entity mention pairs that appear in the same sentence are constructed as intrasentence level instances.
2. If multiple mentions refer to the same entity in a sentence, the mentions in the nearest distance should be constructed as an instance.
3. For instance, chemical D007545 and disease D006332 in *sentence 1* form an intrasentence level positive instance, while chemical D007545 and disease D006984 in *sentence 4* form an intrasentence level negative instance.

Relation Instance Construction for Intersentence Level

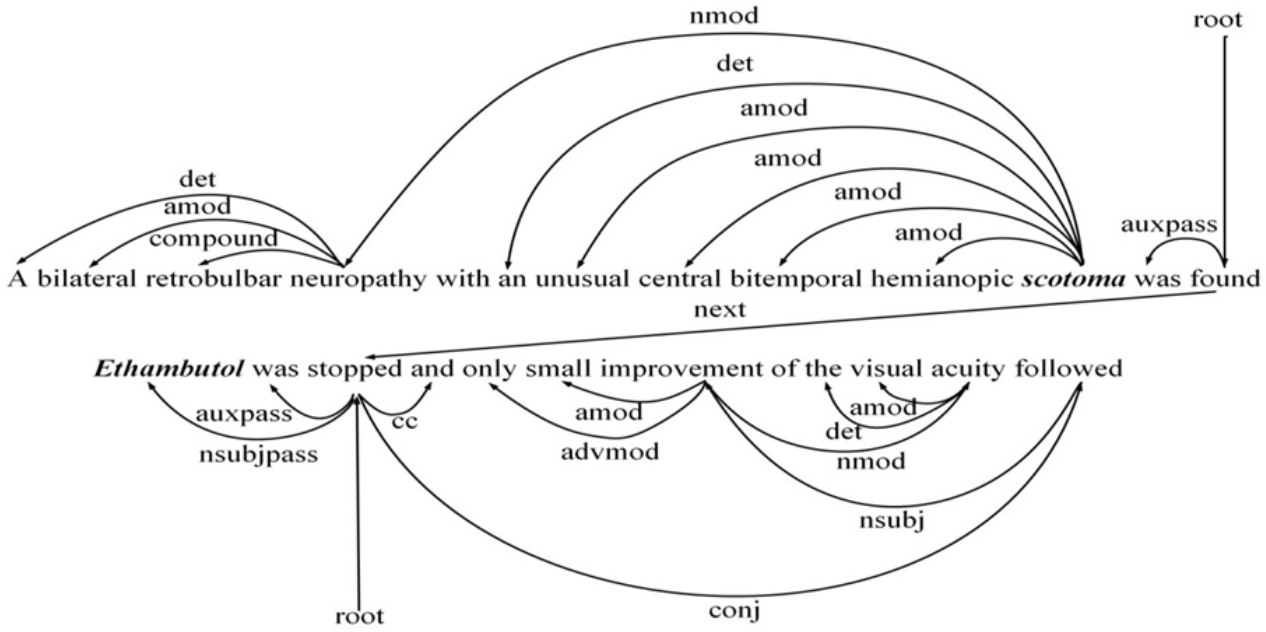
1. Only the chemical-disease entity pairs that are not involved in any intrasentence level are considered as intersentence level instances.
2. If multiple mentions refer to the same entity, the chemical and disease mention in the nearest distance are chosen.

According to our heuristic rules, chemical D007545 in *sentence 4* and disease D006965 in *sentence 5* are regarded as an intersentence level instance because there are no mentions of them in the same sentence. Chemical D007545 in *sentence 1* and disease D006965 in *sentence 5* will be omitted because their distance is not the shortest. Further, chemical D007545 in *sentence 4* and disease D006984 in *sentence 5* are not regarded as an intersentence level instance because chemical D007545 already has intrasentence level co-occurrence with disease D006984 in *sentence 4*.

Document-Level Dependency Graph

To generate features for entity pairs within and across sentences, we introduce a document-level dependency graph with nodes representing words and edges that show intrasentence and intersentence dependency relations. Figure 2 shows an example of document-level dependency graph for 2 sentences. In this study, we use the following 3 types of intrasentence and intersentence dependency edges.

Figure 2. An example of a document-level dependency graph for 2 sentences expressing a CID relation. The chemical and disease entity mention is highlighted in bold. For simplicity, we have omitted self-node edges. CID: chemical-induced disease.

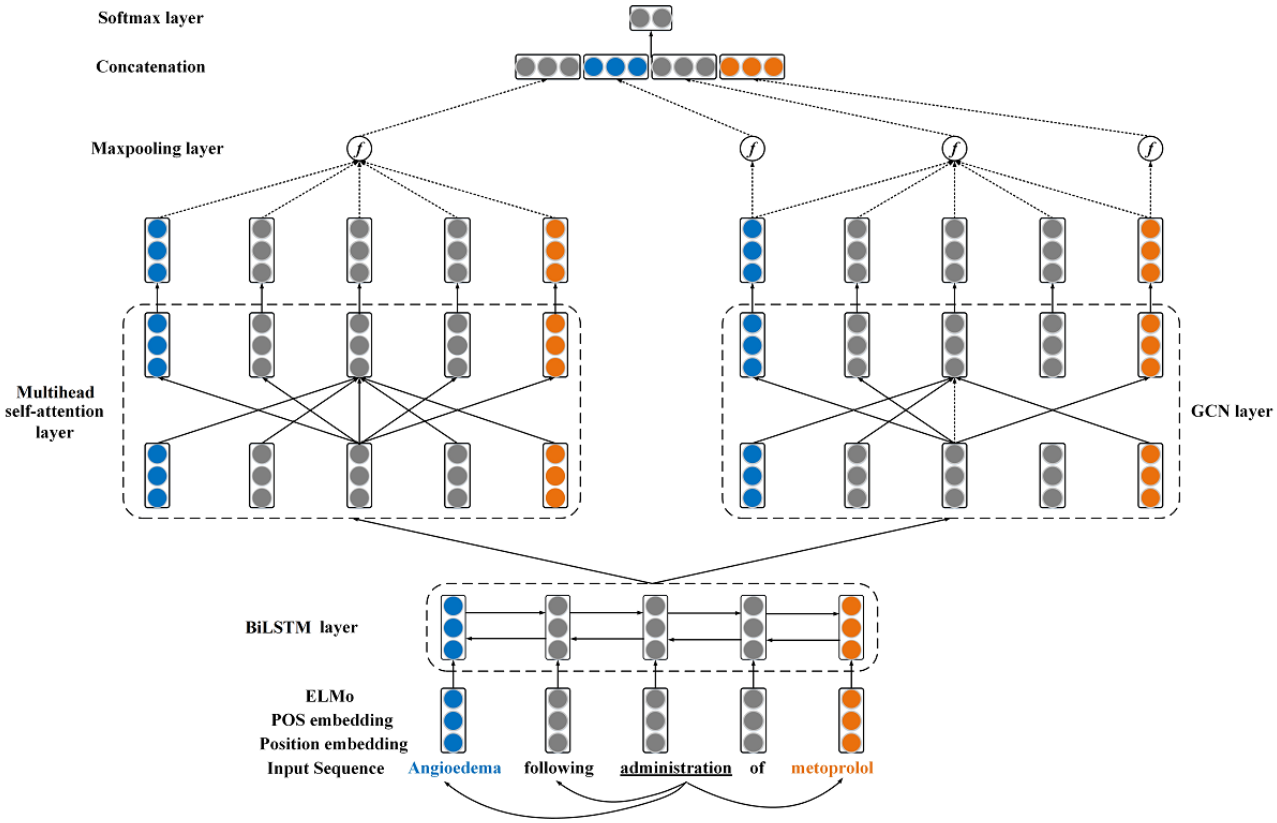


1. Syntactic dependency edge: The syntactic structure is crucial to biomedical relation extraction. Hence, we use syntactic dependency edges derived from Stanford dependency syntactic parser as intrasentential edges. For instance, “conj” denotes the syntactic relation between the word “stopped” and “followed” in the same sentence.
2. Adjacent sentence edge: Dependencies between sentences are useful for document-level relation extraction. Thus, we consider the sentence as a node in a type of discourse dependency tree. Moreover, we added an edge between the dependency roots of adjacent sentences as an intersentential edge, which is a simple but an effective approach. For instance, “next” denotes the syntactic relation between 2 sentences.
3. Self-node edge: We added self-node edges to all the nodes of the graph in order to enable GCN to not only learn information based on neighbor nodes but also learn the node information itself.

Model Architecture

The schematic overview of our model is shown in Figure 3. In short, our model mainly consists of 4 parts: the input representation layer, the BiLSTM layer, the multihead attention layer, and the GCN layer. The inputs of our model are text sequences. The input layer will generate a deep contextualized word representation for each word. Recent studies [22,23] have suggested that the part of speech (POS) and the position of each word are useful for biomedical relation extraction. Hence, we concatenate the deep contextualized word representation and POS and position embedding as the whole word representation. The BiLSTM layer will obtain contextual features from the word representation. The multihead attention layer will apply the self-attention mechanism multiple times to capture the relative semantic features from different representation subspaces. The GCN layer will operate over the document-level dependency graph to capture long-range syntactic features. We employed max pooling over the outputs of the multihead attention layer and the GCN layer and then concatenated these 2 vectors as the final representation. Finally, we employed a fully connected layer and the *Softmax* function to identify the CID relation. Our model will be described in detail in the following section.

Figure 3. Overview of our model. The input representation consists of ELMo, POS embedding, and position embedding. In the multi-head self-attention layer, we only show the detailed self-attention computation for the word "administration." In the GCN layer, we only show the detailed graph convolution computation for the word "administration." BiLSTM: bidirectional long short-term memory; POS: part of speech; GCN: graph convolutional network.



Input Representation

We used ELMo instead of the traditional word representation in our model. Traditional word representation generates a fixed representation vector for the same word. However, ELMo is the function of the entire input sentence based on a bidirectional language model so that it can generate different representation vectors for the same word according to the different sentence context.

Given that a sequence $\{t_1, t_2,\ldots..t_N\}$ denotes the word tokens in a sentence S. Given a token t_k, the forward language model calculates the probability of the token t_k based on the previous tokens $\{t_1, t_{2,\ldots,} t_{(k-1)}\}$ of t_k in the sentence S as follows:

(1)

Similarly, the backward language model calculates the probability of the token t_k based on the back tokens $\{t_1, t_{2, \ldots,} t_{(k-1)}\}$ of t_k in the sentence S as follows:

(2)

Combining the forward and the backward language models as a bidirectional language model, the log-likelihood can be maximized as follows:

(3)

ELMo can represent the semantic and syntactic information of the word. In our model, we use a linear combination of the hidden state in each layer of the bidirectional language model to generate a deep contextualized representation for words. The POS and the position information of a word are crucial to biomedical relation extraction. Therefore, we also utilize POS embedding and position embedding to enhance the representation ability of the input. The POS embedding represents the POS feature of a word, and the position embedding reflects the relative distance between the word and the target entity. Given a word at position i, we obtain its POS embedding $w_{p,i}$ and position embedding $w_{d,i}$ based on mapping matrixes M_p and M_d, respectively. Finally, the whole word representations concatenate deep contextualized word representations, POS embedding, and position embedding as follows:

$$wi=[we,i; wp,i; wd,i] \quad (4)$$

BiLSTM

The LSTM model is a variant of recurrent neural network models that has been used in many NLP tasks successfully. The LSTM model overcomes the vanishing gradient problem by introducing a gating mechanism [24]. Therefore, it is suitable to capture the long-term dependency feature. The LSTM unit consists of 3 components: the input gate i_t, the forget gate f_t, and the output gate o_t. At the time step t, the LSTM unit utilizes the input word x_t, the previous hidden state $h_{(t-1)}$, and the

previous cell state $c_{(t-1)}$ to calculate the current hidden state h_t and cell state c_t. The equations are as follows:

$$ft=\sigma(Wfxt+Ufh(t-1)+bf) \quad (5)$$

$$ot=\sigma(Woxt+Uoh(t-1)+bo) \quad (6)$$

$$gt=tanh(Wgxt+Ugh(t-1)+bg) \quad (7)$$

$$it=\sigma(Wixt+Uih(t-1)+bi) \quad (8)$$

$$c_t=f_t \odot c_{(t-1)}+ i_t \odot g_t \quad (9)$$

$$h_t= o_t \odot tanh(c_t) \quad (10)$$

where W, U, b are the weight and bias parameters, and $\odot$ denotes element-wise multiplication. In this study, we use the BiLSTM model that can capture the forward and backward context features simultaneously. The BiLSTM model combines a forward LSTM and a backward LSTM. Given the hidden state of the forward LSTM  and the hidden state of the backward LSTM  , the final hidden state is concatenated as:

Multihead Attention

The BiLSTM model learns the context features from the input sequences automatically and effectively. However, these features make different contributions to the biomedical relation extraction. In our model, we capture the relatively important features by introducing multihead attention mechanism. The essence of multihead attention is applying self-attention mechanism multiple times so that it may let the model learn the relatively important features from different representation subspaces. The self-attention mechanism generates the output based on a query and a set of key-value pairs. The output is the weighted sum of the values, where the weight assigned to each value is computed by applying attention function to the query with the corresponding key. In our study, we deal with the output of the BiLSTM model by multihead self-attention. Further, we use the dot-product attention function instead of the standard additive attention function [25] as follows:

(11),

where Q, K, $V \in R^n$ represent query, key, and value matrixes, respectively. d is the dimension of the output of the BiLSTM model.

The main idea of the multihead attention is applying the self-attention mechanism multiple times. If the multihead attention contains h heads, the i-th attention head can be calculated as $head_i=Attention\ (Q_i,K_i,V_i)$. Thus, the final multihead attention is the concatenation of $\{head_1,head_2,...,head_h\}$ as $MultiHead\ (Q,K,V)=Concat\ (head_1,head_2,...,head_h)\ W^o$. The output of the multihead attention layer is a matrix of R^{nat}.

GCN

GCN is an adaptation of CNN [26], which operates on graphs. Given a graph with n nodes, the graph structure can be represented as an adjacency matrix A. In this study, we converted the document-level dependency graph into its corresponding adjacency matrix A, where $A_{ij}=1$ if there is a dependency edge going from token i to token j; otherwise $A_{ij}=0$. The dependency graph can be calculated as an undirected graph [27], which means $A_{ij}=A_{ji}$. Further, we add a self-node edge to all the nodes in the graph, which means $A_{ii}=1$. Since the degree of a node in the dependency graph varies a lot, this may bias the output representation toward favoring high-degree nodes, regardless of the information carried in the node. To solve this issue, we normalize the activations in the graph convolution before feeding it through the nonlinearity. Finally, the graph convolution operation for node i at the l-th layer where  and  denote the input representation and the output representation of node can be defined as follows:

(12),

where $W^{(l)}$ is the weight matrix, $b^{(l)}$ is the bias vector,

is the degree of node i in the dependency graph, and ρ is an activation function (eg, a rectified linear unit).

The GCN model takes the output of the BiLSTM model as the input word representation:

Then, we stack the graph convolution operation over layers and obtain

as the output word representations of the GCN model. Note that the GCN model presented above uses the same parameters for all edges in the dependency graph.

Relation Classification

To make use of the output word representation of the GCN model for relation extraction, we generate the sentence representation as follows:

$$h_{sent}=f\ (h^{(L)})=f\ (GCN(h^{(0)}) \quad (13)$$

where $h^{(L)}$ denotes the output representations at the last layer L of the GCN model, and $f:R^{n:R} \rightarrow R^d$ is a max-pooling function that maps n output vectors to the sentence vector.

Inspired by recent studies [28,29], entity information is central to relation classification. Therefore, we also obtain the chemical entity representation h_c as shown in  . Similarly, we can obtain the disease entity representation h_d. The feature representation of the whole GCN model is $h_{GCN}=[h_{sent}; h_c; h_d]$.

We also obtain the feature representation h_{att} from the output of the multihead attention layer by applying max pooling to the multihead attention matrix. We concatenate h_{GCN} and h_{att} to form the final representation $h_{final}=[h_{GCN}; h_{att}]$ for relation classification. Then, the final representation is fed into a 2-layer perceptron as follows:

(14) and (15).

Finally, the hidden representation h_2 is fed to a *Softmax* function to calculate the confidence of the CID relation:

$$o=softmax\ (W_o h_2+b_o)\ (16)$$

where o is the output, W_o is the weight matrix, and b_o is the bias vector.

Results

Dataset

We evaluated our model on the CDR corpus, which was released by the BioCreative V task. The CDR dataset is the benchmark dataset for the CID relation extraction task, which consists of 1500 PubMed abstracts—500 each for training, development, and test set. Table 1 shows the details of the dataset.

Table 1. Statistics of the chemical-disease relation dataset.

Task dataset	Abstracts (n=1500)	Chemical-induced disease relations (n=3116)
Training	500	1038
Development	500	1012
Test	500	1066

In this study, the gold entity annotations provided by BioCreative V were used to evaluate our model. All the comparison methods reported in this paper were evaluated with gold entity annotations. Therefore, it is fair and comparable. Further, we measured the CID relation extraction performance with precision, recall, and F-measure.

Experimental Settings

The dimensions of POS embedding and position embedding are both 100. The dimension of ELMo is 1024. The dimensions of the LSTM hidden layer and the GCN layer are 500 with the dropout proportion of 0.5. The dimensions of 2-layer perceptron are also 500 with the dropout proportion of 0.5. Our model was trained by Adam [30] with a learning rate of 0.001 and a minibatch size of 32. In addition, our model was implemented based on an open-source deep learning library PyTorch [31]. We used StanfordNLP [32] to obtain the POS of the word and the dependency tree. Further, we used the pretrained ELMo representations for the deep contextualized word representations.

Experimental Results

Effect of Input Representation

We evaluated the effectiveness of the input representation of our model. We used the same model that we proposed and changed the input representations. The comparison performance of the different input representations is presented in Table 2.

Table 2. The effect of the input representation on performance.

Input representation	Precision (%)	Recall (%)	F-measure (%)
Word[a]	47.3	71.7	57.0
Word+position[b]	49.1	71.4	58.2
Word+position+POS[c]	51.6	71.8	*60.1*
ELMo[d]	57.0	67.4	61.8
ELMo+position[e]	54.2	74.9	62.9
ELMo+position+POS[f]	56.3	72.7	*63.5*
BioBERT+position+POS[g]	57.9	70.1	63.4

[a]The input representation of the model is the word embedding, which is pretrained by word2vec.

[b]The input representation of the model is the concatenation of the word embedding and position embedding.

[c]The input representation of the model is the concatenation of the word embedding, position embedding, and part of speech (POS) embedding. The F-measure (%) for this representation was an important finding.

[d]The input representation of the model is the deep contextualized word representation.

[e]The input representation of the model is the deep contextualized word representation and position embedding.

[f]The input representation of the model is the deep contextualized word representation, position embedding, and POS embedding. The F-measure (%) for this representation was an important finding.

[g]The word representation is generated from the last hidden layer of the bidirectional encoder representations from transformers for biomedical text mining (BioBERT) [33] in a feature-based approach, which means that the parameters of the BioBERT are not fine-tuned. The input representation of the model is the BioBERT word representation, position embedding, and POS embedding.

In Table 2, we can observe that the model achieves an F-measure of 57.0% when we only use the pretrained word embedding as the input representation. When we concatenate the pretrained word embedding and position embedding, the F-measure is

improved from 57.0% to 58.2%, which yields a 1.2% improvement. When we concatenate the pretrained word embedding, position embedding, and POS embedding as the input representations, we yield another 1.9% improvement compared with only using the pretrained word embedding and position embedding. The result indicates that both POS and position features are effective for the CID relation extraction. The deep contextualized word representation ELMo significantly outperforms the pretrained word embedding and yields a 4.8% improvement in the F-measure. The result indicates that ELMo can generate a more comprehensive representation for the word according to the sentence context, which results in a better CDR performance. Similarly, combining the position and POS embedding with the deep contextualized word representation can further improve the performance. When we concatenate the deep contextualized word representation, position embedding, and POS embedding as the input representation, we achieve the best F-measure of 63.5%. We also use the word representations generated from the bidirectional encoder representations from transformers for biomedical text mining in a feature-based approach and achieve an F-measure of 63.4%, which is similar to using ELMo.

Effect of the Attention Mechanism

We evaluated the effectiveness of the multihead self-attention mechanism. We used the same model architecture that we proposed, but we dealt with the output of BiLSTM by different attention mechanisms. The attention mechanism is divided into 2 categories: single-head attention mechanism and multihead attention mechanism. In single-head attention mechanism, we use 3 types of attention function: additive attention, general attention, and scaled dot-product attention, as shown below.

(17)

(18)

(19)

where h_i is the output of the BiLSTM, W_1, W_2, s, v are the parameter matrixes, and d is the dimension of the output of the BiLSTM model. The formula of the multihead attention is described in formula (11). The comparison performance of the different attention mechanism is presented in Table 3.

Table 3. The effect of the attention mechanism on performance.

Attention mechanism	Precision (%)	Recall (%)	F-measure (%)
Without attention	55.1	71.3	62.2
Additive attention	55.9	70.3	62.3
General attention	55.3	71.8	62.5
Scaled dot-product attention	54.9	73.3	62.8
Multihead attention	56.3	72.7	63.5

In Table 3, we can see that using the attention mechanism can improve the performance of the CID relation extraction. The multihead attention mechanism is more helpful than other single-head attention mechanisms. This suggests that the multihead attention mechanism can capture more valuable features from different representation subspaces.

Effect of the Attention Heads

We evaluated the effectiveness of the number of heads of the multihead attention mechanism. In this comparative experiment, we used the deep contextualized word representation, position embedding, and POS embedding as the input representation, and the dimensions of query, key, and value are the same. As shown in Table 4, we only varied the number of heads of the multihead attention.

Table 4. The effect of the attention heads on performance.

Heads (n)	Precision (%)	Recall (%)	F-measure (%)
2	57.2	68.2	62.2
4	56.9	70.6	63.0
5	56.3	72.7	63.5
8	57.0	70.2	62.9
10	54.4	75.4	63.2

In Table 4, we can see that the multihead attention mechanism can effectively improve the performance of the CID relation extraction. We can observe that the F-measure ranges from 62.2% to 63.5% when setting a different number of heads. When the number of heads is too little or too large, the performance will drop off. In short, we achieve the best F-measure of 63.5% when we set the number of heads as 5.

Ablation Study

To examine the contributions of the 2 main components, namely, multihead attention layer and GCN layer, we ran an ablation study. The experimental results are shown in Table 5. The results contain intrasentence level, intersentence level, and relation merging, which means that merging the intrasentence and intersentence level results in the final document-level result.

Table 5. An ablation study for our model.a

Model	Intrasentence level			Intersentence level			Relation merging		
	Precision (%)	Recall (%)	F-measure (%)	Precision (%)	Recall (%)	F-measure (%)	Precision (%)	Recall (%)	F-measure (%)
Without multi-head attention	58.2	82.9	68.4	44.7	44.3	44.5	55.1	71.3	62.2
Without GCN[b]	62.6	74.1	67.9	43.6	48.4	45.9	57.1	66.4	61.4
Our model	59.1	81.5	*68.5*	47.8	52.2	*49.9*	56.3	72.7	*63.5*

[a]The values in italics indicate significant findings.

[b]GCN: graph convolutional network.

We can observe that removing either the multihead attention layer or the GCN layer reduces the performance of the model. This suggests that both layers can learn effective features. When we remove the multihead attention layer and the GCN layer, the F-measure drops by 1.3% and 2.1%, respectively. In particular, we can observe that adding either the multihead attention layer or the GCN layer improves the performance in the intersentence level relation extraction by a large margin. When we remove the multihead attention layer and the GCN layer, the intersentence level F-measure drops by 5.4% and 4.0%, respectively. This suggests that the multihead attention layer can capture the relatively important features from different representation subspaces and the GCN layer can capture long-range syntactic features for intersentence level relation extraction.

Comparison with Related Work

We compared our model with several state-of-the-art methods of the CID relation extraction. These methods are divided into 2 categories: methods without additional resources (without knowledge bases) and methods using additional resources (with knowledge bases). These following methods have been summarized in Table 6.

1. Pattern rule-based: Lowe et al [5] developed a pattern-based system with some heuristic rules to extract CID relations within the same sentence, and they achieved an F-measure of 60.8%.
2. Maximum entropy model: Gu et al [7] developed a machine learning-based system that utilized simple but effective manual linguistic features with the maximum entropy model. They built rich manual features for intrasentence level and intersentence level instances. They achieved an F-measure of 58.3%.
3. LSTM+ support vector machine (SVM): Zhou et al [10] developed a hybrid system, which consists of a feature-based model that utilized flat features and structure features with SVM and a neural network model based on LSTM. Their model achieved an F-measure of 56.0%. After using additional postprocessing heuristic rules, they achieved a 5.3% improvement in the F-measure.
4. CNN+maximum entropy: Gu et al [11] proposed a maximum entropy model for intersentence level relation extraction and a CNN model for intrasentence level relation extraction. They achieved an F-measure of 60.2%. They also used additional postprocessing heuristic rules to improve performance that increases the F-measure to 61.3%.
5. Biaffine Relation Attention Network: Verga et al [12] proposed this based on the multihead self-attention model, which can predict relationships between all the mentioned pairs in the document. The model achieved an F-measure of 62.1%.
6. Graph convolutional neural network: Sahu et al [18] proposed a labelled edge graph convolutional neural network model on a document-level graph. The model achieved an F-measure of 58.6%.
7. SVM_Xu: Xu et al [6] explored 4 different knowledge bases to extract the knowledge features and achieved an F-measure of 67.2%.
8. SVM_Pons: Pons et al [9] extracted 3 sets of features, which are prior knowledge and statistical and linguistic information from the document. They achieved an F-measure of 70.2%.
9. Knowledge-guided convolutional network: Zhou et al [34] proposed a CNN that integrated both relation representations and entity representations learned from knowledge bases. The model achieved an F-measure of 71.3%.

Table 6. Comparisons with related work.

Category and method	Precision (%)	Recall (%)	F-measure (%)
Without knowledge bases			
Lowe et al [5]			
Pattern rule-based	59.3	62.3	60.8
Gu et al [7]			
ME[a]	62.0	55.1	58.3
Zhou et al [10]			
LSTM+SVM[b]	64.9	49.3	56.0
LSTM+SVM+PP[c]	55.6	68.4	61.3
Gu et al [11]			
CNN+ME[d]	60.9	59.5	60.2
CNN+ME+PP	55.7	68.1	61.3
Verga et al [12]			
BRAN[e]	55.6	70.8	62.1
Sahu et al [18]			
GCNN[f]	52.8	66.0	58.6
Our study			
GCN[g]+Multihead attention	56.3	72.7	63.5
With knowledge bases			
Xu et al [6]			
SVM	65.8	68.6	67.2
Pons et al [9]			
SVM	73.1	67.6	70.2
Zhou et al [34]			
KCN[h]	69.7	72.9	71.3

[a]ME: maximum entropy model.

[b]LSTM+SVM: long short-term memory+support vector machine.

[c]LSTM+SVM+PP: long short-term memory+support vector machine+postprocessing.

[d]CNN+ME: convolutional neural network+maximum entropy model.

[e]BRAN: biaffine relation attention network.

[f]GCNN: graph convolutional neural network.

[g]GCN: graph convolutional network.

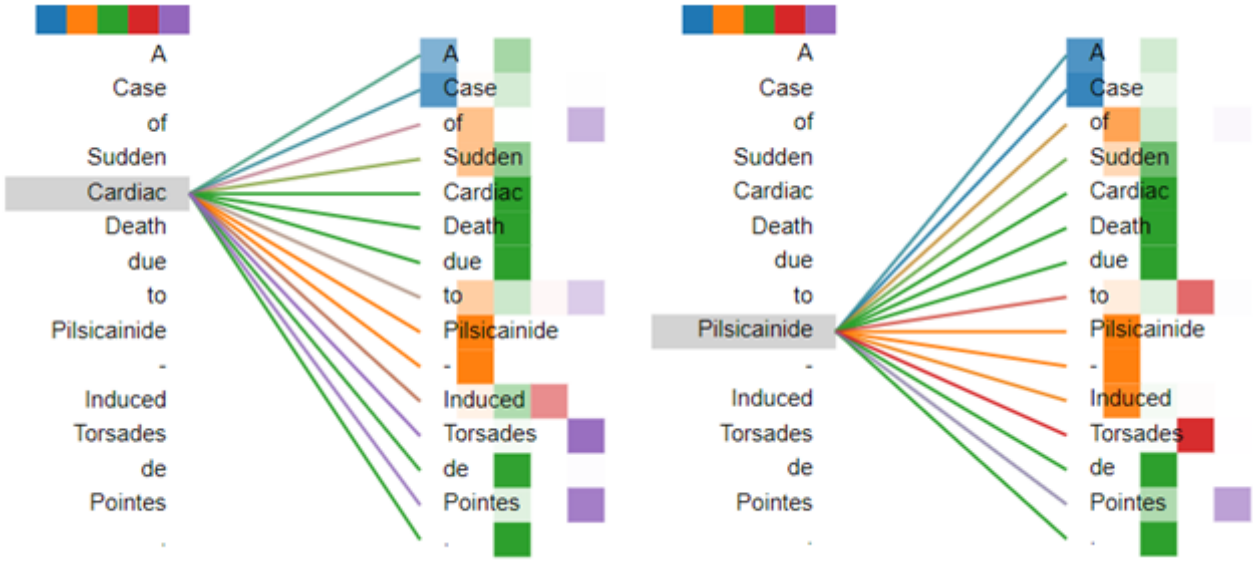
[h]KCN: knowledge-guided convolutional networks.

In Table 6, the deep neural network-based methods achieved competitive performance in the CID relation extraction task. For example, Sahu et al [18] used GCN to capture dependency information and achieved an F-measure of 58.6%. Compared with other deep neural network-based methods, we not only employed the multihead attention to capture the relatively important semantic features but also used the GCN to capture the valuable syntactic features from the document-level dependency graph automatically and effectively. We also observed that some studies [7,10,11] designed and extracted rich semantic and syntactic features for the relation extraction task and used additional postprocessing heuristic rules to improve performance. Our method is an end-to-end neural network-based model and achieves a high F-measure of 63.5% without using postprocessing heuristic rules. As shown in Table 6, the methods with knowledge bases outperform the methods without knowledge bases significantly. This suggests that prior knowledge is much useful for CID relation extraction. In this study, we focus on the effectiveness of GCN and multihead attention mechanism rather than the prior knowledge. We will attempt to integrate the biomedical knowledge to further improve the performance of our method in our future work.

Visualization of Multihead Attention Mechanisms

To understand our multihead self-attention mechanism clearly, we visualized the attention weights of an example sequence in Figure 4. Different colors represent different heads. The darker the color is, the higher the attention weight is. In Figure 4, the word pays different levels of attention to different words in different heads. For the word “Cardiac,” the word “Pilsicainide” has the higher weight score in the second head; however, the words “Torsades” and “Pointes” have the higher weight score in the last head. For the word “Pilsicainide,” the words “Cardiac” and “Death” have the higher weight score in the third head; however, the word “Torsades” has the higher weight score in the fourth head. Thus, the multihead self-attention mechanism can make the model capture the relatively important features from different representation subspaces.

Figure 4. Examples of the multi-head self-attention mechanism. Attentions here shown only for the words "Cardiac" and "Pilsicainide." Different colors represent different heads.



Error Analysis

To understand our model better, we performed an error analysis on the output of our final results. There are the 2 main types of errors: false positive errors and false negative errors. We list some examples to analyze the errors. In false positive errors, some instances are nonrelations but are mistaken as CID relations. For the sentences “Carbamazepine (Chemical: D002220)-induced cardiac dysfunction (Disease: D006331)” and “A patient with sinus bradycardia and atrioventricular block (Disease: D054537) induced by carbamazepine (Chemical: D002220),” the disease D006331 is the hypernym of the disease D054537. According to the labeling rules of the CDR corpus, we need to extract the most specific relations. Thus, the first sentence does not express a CID relation and the second sentence expresses a CID relation. However, our model extracts a CID relation between the chemical D002220 and the disease D006331 in the first sentence incorrectly because the first sentence is the common sentence pattern that expresses a CID relation. In false negative errors, several CID relations are not recognized. One of the main reasons for some intersentence level instances to be removed by the heuristic rules in the relation instance construction stage is because the sentence distance is more than 3. In the future, we will consider preferable preprocessing and postprocessing techniques to solve the above problems.

Discussion

In this paper, we propose a novel end-to-end neural network based on GCN and multihead attention. The document-level dependency graph is constructed to capture the dependency syntactic information across sentences. We applied GCN to capture the long-range dependency syntactic features, which can improve the performance of intersentence level relation extraction. Further, we employed the multihead attention mechanism to capture the relatively important context features from different semantic subspaces. ELMo is used in our model to enhance the input representation. We evaluate the effectiveness of ELMo, multihead attention mechanism, and GCN on the BioCreative V CDR dataset. Experimental results show that ELMo, multihead attention, and GCN can significantly improve the performance of the CDR extraction. Our method achieves an F-measure of 63.5%, which is superior to other state-of-the-art methods. There are many large-scale knowledge bases such as the Comparative Toxicogenomics Database, Unified Medical Language System, Medical Subject Headings, UniProt, and the commercial system Euretos Knowledge Platform. These knowledge bases contain a large amount of structured data in the form of triples (entity, relation, entity), wherein relation represents the relationship between 2 entities. Some studies suggest that integrating the structured information from the knowledge bases may improve the performance of the CDR extraction. In future studies, we will

integrate the biomedical knowledge to further improve the performance of our method.

Acknowledgments

The work was supported by grants from National Natural Science Foundation of China (No. 61572098 and 61572102). We would like to thank the Natural Science Foundation of China. We also would like to thank all the anonymous reviewers for their valuable suggestions and constructive comments.

Authors' Contributions

JW and YZ led the method application, experiment conduction, and the result analysis. XC, YZ, and JW participated in the data extraction and preprocessing. YZ and XW participated in the manuscript revision. HL and ZY provided theoretical guidance and the revision of this paper.

Conflicts of Interest

None declared.

References

1. Islamaj Dogan R, Murray GC, Névéol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database (Oxford) 2009 Nov 27;2009:bap018-bap018 [FREE Full text] [doi: 10.1093/database/bap018] [Medline: 20157491]
2. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 2013 Apr 15;93(4):335-341. [doi: 10.1038/clpt.2013.1] [Medline: 23443757]
3. Qu J, Ouyang D, Hua W, Ye Y, Li X. Distant supervision for neural relation extraction integrated with word attention and property features. Neural Netw 2018 Apr;100:59-69. [doi: 10.1016/j.neunet.2018.01.006] [Medline: 29471196]
4. Wei C, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford) 2016;2016:baw032 [FREE Full text] [doi: 10.1093/database/baw032] [Medline: 26994911]
5. Lowe DM, O'Boyle NM, Sayle RA. Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall. Database (Oxford) 2016 Apr 08;2016:baw039 [FREE Full text] [doi: 10.1093/database/baw039] [Medline: 27060160]
6. Xu J, Wu Y, Zhang Y, Wang J, Lee HJ, Xu H. CD-REST: a system for extracting chemical-induced disease relation in literature. Database (Oxford) 2016;2016:baw036 [FREE Full text] [doi: 10.1093/database/baw036] [Medline: 27016700]
7. Gu J, Qian L, Zhou G. Chemical-induced disease relation extraction with various linguistic features. Database (Oxford) 2016 Apr 06;2016:baw042 [FREE Full text] [doi: 10.1093/database/baw042] [Medline: 27052618]
8. Zhou H, Deng H, He J. Chemical-disease relations extraction based on the shortest dependency path tree. 2015 Presented at: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; 2015; Seville, Spain p. 214-219 URL: https://pdfs.semanticscholar.org/e66a/754947a9abd6665ab16815f52bc1c9aed596.pdf
9. Pons E, Becker BF, Akhondi SA, Afzal Z, van Mulligen EM, Kors JA. Extraction of chemical-induced diseases using prior knowledge and textual information. Database (Oxford) 2016 Apr 14;2016:baw046 [FREE Full text] [doi: 10.1093/database/baw046] [Medline: 27081155]
10. Zhou H, Deng H, Chen L, Yang Y, Jia C, Huang D. Exploiting syntactic and semantics information for chemical-disease relation extraction. Database (Oxford) 2016 Apr 14;2016:baw048 [FREE Full text] [doi: 10.1093/database/baw048] [Medline: 27081156]
11. Gu J, Sun F, Qian L, Zhou G. Chemical-induced disease relation extraction via convolutional neural network. Database (Oxford) 2017 Jan 01;2017(1):bax024 [FREE Full text] [doi: 10.1093/database/bax024] [Medline: 28415073]
12. Verga P, Strubell E, Mccallum A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. 2018 Presented at: the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018; New Orleans, USA. [doi: 10.18653/v1/N18-1080]
13. Nguyen D, Verspoor K. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. 2018 Presented at: Proceedings of the BioNLP workshop; 2018; Melbourne, Australia p. 129-136. [doi: 10.18653/v1/W18-2314]
14. Peters ME, Neumann M, Iyyer M. Deep contextualized word representations. 2018 Presented at: the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018; New Orleans, USA p. 2227-2237. [doi: 10.18653/v1/N18-1202]
15. Li L, Nie Y, Han W, Huang J. A Multi-attention-Based Bidirectional Long Short-Term Memory Network for Relation Extraction. 2017 Presented at: International Conference on Neural Information Processing; 2017; Guangzhou, China p. 216-227. [doi: 10.1007/978-3-319-70139-4_22]

16. Vaswani A, Shazeer N, Parmar N. Attention is all you need. 2017 Presented at: Neural Information Processing Systems(NIPS); 2017; Long Beach, USA p. 5998-6008 URL: https://dl.acm.org/doi/pdf/10.5555/3295222.3295349
17. Quirk C, Poon H. Distant supervision for relation extraction beyond the sentence boundary. 2017 Presented at: the 15th Conference of the European Chapter of the Association for Computational Linguistics; 2017; Valencia, Spain p. 1171-1182. [doi: 10.18653/v1/e17-1110]
18. Sahu SK, Christopoulou F, Miwa M, Ananiadou S. Inter-sentence Relation Extraction with Document-level Graph Convolutional Neural Network. 2019 Presented at: the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Florence, Italy p. 4309-4316. [doi: 10.18653/v1/p19-1423]
19. Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-Aryamontri A, Winter A, et al. The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinformatics 2011 Oct 03;12 Suppl 8:S3 [FREE Full text] [doi: 10.1186/1471-2105-12-S8-S3] [Medline: 22151929]
20. Segura-Bedmar I, Martínez P, Herrero-Zazo M. Lessons learnt from the DDIExtraction-2013 Shared Task. J Biomed Inform 2014 Oct;51:152-164 [FREE Full text] [doi: 10.1016/j.jbi.2014.05.007] [Medline: 24858490]
21. Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. J Am Med Inform Assoc 2001;8(4):317-323 [FREE Full text] [doi: 10.1136/jamia.2001.0080317] [Medline: 11418538]
22. Zhao Z, Yang Z, Luo L, Lin H, Wang J. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics 2016 Nov 15;32(22):3444-3453 [FREE Full text] [doi: 10.1093/bioinformatics/btw486] [Medline: 27466626]
23. Zhang Y, Zheng W, Lin H, Wang J, Yang Z, Dumontier M. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics 2018 Mar 01;34(5):828-835 [FREE Full text] [doi: 10.1093/bioinformatics/btx659] [Medline: 29077847]
24. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
25. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2015 Presented at: 3rd International Conference on Learning Representations; May 7-9, 2015; San Diego, CA, USA URL: https://arxiv.org/pdf/1409.0473.pdf
26. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. IEEE 1998 Nov;86(11):2278-2324. [doi: 10.1109/5.726791]
27. Zhang Y, Qi P, Manning CD. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. 2018 Presented at: Conference on Empirical Methods in Natural Language Processing; 2018; Brussels, Belgium p. 2205-2215. [doi: 10.18653/v1/d18-1244]
28. Santoro A, Raposo D, Barrett D, Malinowski M, Pascanu R, Battaglia P, et al. A simple neural network module for relational reasoning. 2017 Presented at: Advances in Neural Information Processing Systems; 2017; Long Beach, USA p. 4967-4976 URL: https://papers.nips.cc/paper/7082-a-simple-neural-network-module-for-relational-reasoning.pdf
29. Lee K, He L, Lewis M, Zettlemoyer L. End-to-end neural coreference resolution. 2017 Presented at: the 2017 Conference on Empirical Methods in Natural Language Processing; 2017; Copenhagen, Denmark p. 188-197. [doi: 10.18653/v1/d17-1018]
30. Kingma DP, Ba J. Adam:A Method for Stochastic Optimization. 2015 Presented at: the 3rd International Conference for Learning Representations; 2015; San Diego URL: https://arxiv.org/pdf/1412.6980.pdf
31. An open source machine learning framework that accelerates the path from research prototyping to production deployment. PyTorch: From Research to Production. URL: https://pytorch.org/ [accessed 2020-04-08]
32. Software. The Stanford Natural Language Processing Group. URL: https://nlp.stanford.edu/software/ [accessed 2020-05-04]
33. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
34. Zhou H, Lang C, Liu Z, Ning S, Lin Y, Du L. Knowledge-guided convolutional networks for chemical-disease relation extraction. BMC Bioinformatics 2019 May 21;20(1):260 [FREE Full text] [doi: 10.1186/s12859-019-2873-7] [Medline: 31113357]

Abbreviations

BiLSTM: bidirectional long short-term memory
CDR: chemical-disease relation
CID: chemical-induced disease
CNN: convolutional neural network
GCN: graph convolutional network
LSTM: long short-term memory
NLP: natural language processing
POS: part of speech
SVM: support vector machine

Edited by T Hao, B Tang, Z Huang; submitted 30.12.19; peer-reviewed by I Gabashvili, L Li; comments to author 01.03.20; revised version received 14.04.20; accepted 25.04.20; published 31.07.20.

Please cite as:
Wang J, Chen X, Zhang Y, Zhang Y, Wen J, Lin H, Yang Z, Wang X
Document-Level Biomedical Relation Extraction Using Graph Convolutional Network and Multihead Attention: Algorithm Development and Validation
JMIR Med Inform 2020;8(7):e17638
URL: https://medinform.jmir.org/2020/7/e17638
doi:10.2196/17638
PMID:32459636

Original Paper

An Artificial Intelligence Fusion Model for Cardiac Emergency Decision Making: Application and Robustness Analysis

Liheng Gong[1], MD; Xiao Zhang[1], MD; Ling Li[1], MD

Hebei North University, School of Information Science and Engineering, Zhangjiakou, China

Corresponding Author:
Xiao Zhang, MD
Hebei North University
School of Information Science and Engineering
No. 11 Zuanshi South Road
Zhangjiakou,
China
Phone: 86 03134029808
Email: r78z@foxmail.com

Abstract

Background: During cardiac emergency medical treatment, reducing the incidence of avoidable adverse events, ensuring the safety of patients, and generally improving the quality and efficiency of medical treatment have been important research topics in theoretical and practical circles.

Objective: This paper examines the robustness of the decision-making reasoning process from the overall perspective of the cardiac emergency medical system.

Methods: The principle of robustness was introduced into our study on the quality and efficiency of cardiac emergency decision making. We propose the concept of robustness for complex medical decision making by targeting the problem of low reasoning efficiency and accuracy in cardiac emergency decision making. The key bottlenecks such as anti-interference capability, fault tolerance, and redundancy were studied. The rules of knowledge acquisition and transfer in the decision-making process were systematically analyzed to reveal the core role of knowledge reasoning.

Results: The robustness threshold method was adopted to construct the robustness criteria group of the system, and the fusion and coordination mechanism was realized through information entropy, information gain, and mutual information methods.

Conclusions: A set of fusion models and robust threshold methods such as the R2CMIFS (treatment mode of fibroblastic sarcoma) model and the RTCRF (clinical trial observation mode) model were proposed. Our study enriches the theoretical research on robustness in this field.

(JMIR Med Inform 2020;8(7):e19428) doi:10.2196/19428

KEYWORDS

artificial intelligence; fusion model; cardiac emergency; robustness

Introduction

Background

Based on data released by the organizing committee of the 15th annual meeting of the Asian Society of Cardiovascular Surgery, we estimate there are currently 60 million potential heart disease patients in China, of which 8 million patients require cardiovascular surgery. We also surmise that globally two-thirds of heart attack patients can recover, but if ambulance service is not delivered in a timely or improper manner, the proportion of deaths will rise dramatically. In the United States, where medical resources are relatively abundant, the Institute of Medical Research's study of medical institutions in the past 40 years has demonstrated that 7% of patients have suffered as a result of serious medical errors. The World Health Organization reports that in countries such as Canada, New Zealand, and the United Kingdom, 10% of patients suffer from one medical adverse event every year [1]. The accuracy of the doctor's diagnosis, the implementation of ambulance service to patients, and a low error rate must be pursued. However, various errors frequently occur during medical treatment, which pose a threat to the life of patients [2].

As a country with relatively scarce medical resources, there are few theoretical and practical studies in the fields of medical

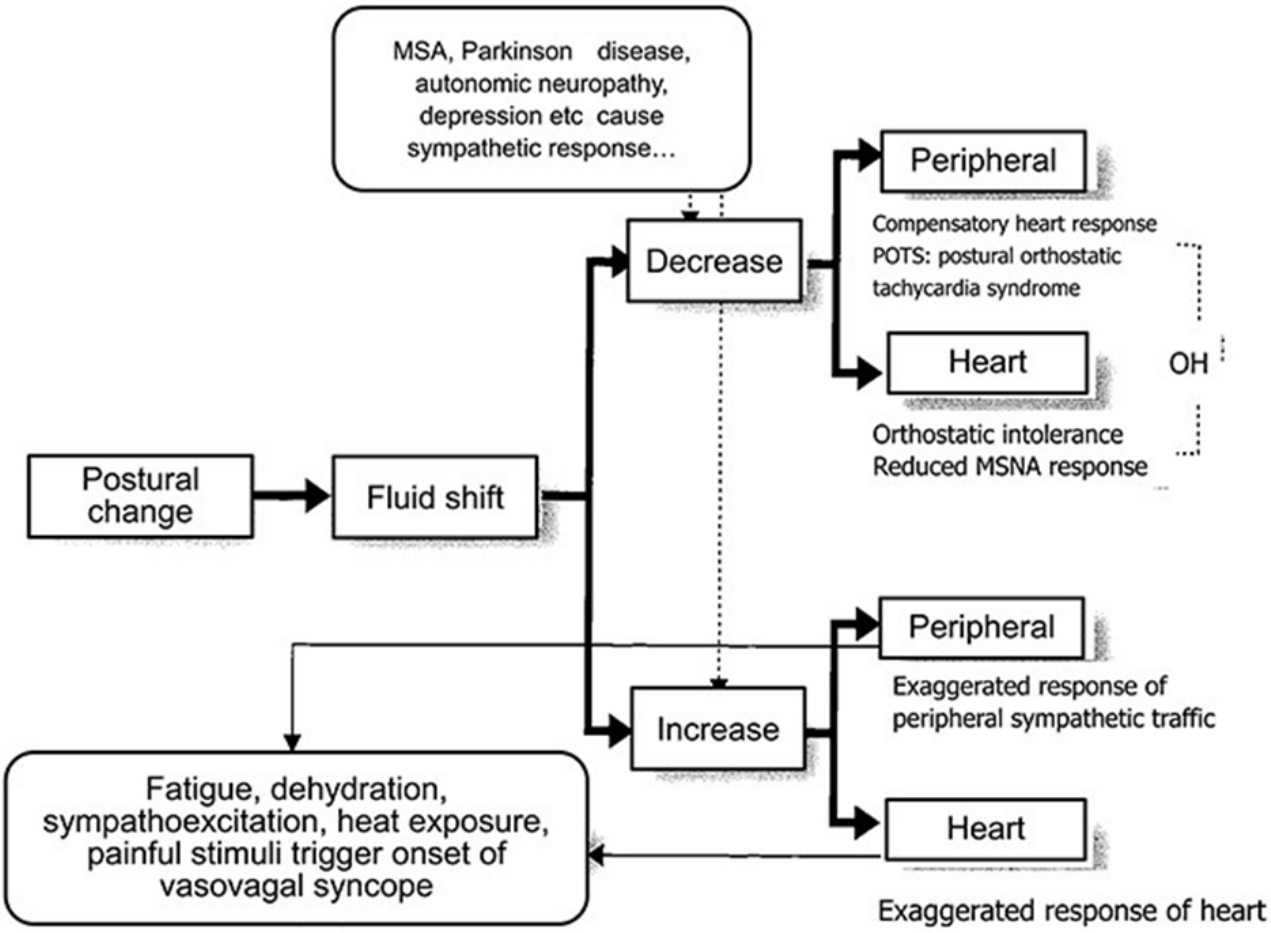
quality and efficiency in China. The existing research focuses primarily on hospital management and the medical expert system. Current research on improving medical quality and reducing adverse events during medical treatment is typically based on the experiences of countries and regions with advanced medical and health management, such as the United States, Singapore, Taiwan, and Hong Kong. From the perspective of management science, the hospital is regarded as a service industry. From the perspective of patients, it is important to strengthen hospital management, improve the level and quality of physical therapy services, and reduce medical costs [3]. With the robustness of medical emergency decision making as the starting point, research on how to improve the quality and efficiency of the medical system is scarce. Existing research focuses mainly on the attitudes of ambulance personnel in the operating room and the development of personnel training models. From a medical perspective, the current diagnosis and treatment of cardiovascular diseases focus primarily on the development of medical technologies for specific diagnoses and treatments, while neglecting the study of models and systems, with even less research on medical process control.

The State Space of Cardiac Emergency Decision-Making Information and Its Description

The CMO-EMPHA First Aid Process

According to the CMO-EMPHA knowledge framework system, the basic cardiac emergency protocol includes nine major first-aid procedures: (1) rescue procedures and analysis of cardiac arrest, (2) cardioversion procedures and analysis, (3) emergency cardiac procedures and analysis, (4) a mechanical separation processing program and analysis, (5) a cardiac pacing program and analysis, (6) a bradycardia processing program and analysis, (7) an acute myocardial infarction rescue program and analysis, (8) a tachycardia processing program and analysis, and (9) first aid for ventricular (VT) fibrillation and VT procedures and analysis. In order to describe the process of cardiac emergency experts' observations, emergency decision making, and treatment of patients during the emergency decision-making process, Figure 1 shows the workflow description of the nine procedures mentioned above. The main operating nodes of this process include tracheal intubation, opening the airway, oxygen inhalation, suction, carotid pulsation examination, anterior cardiac stroke, cardiopulmonary resuscitation (CPR), defibrillation, pacing, and medication.

Figure 1. The CMO-EMPHA emergency cardiac rescue process. MSA: multiple system atrophy; MSNA: muscle sympathetic nerve activity; OH: orthostatic hypotension.

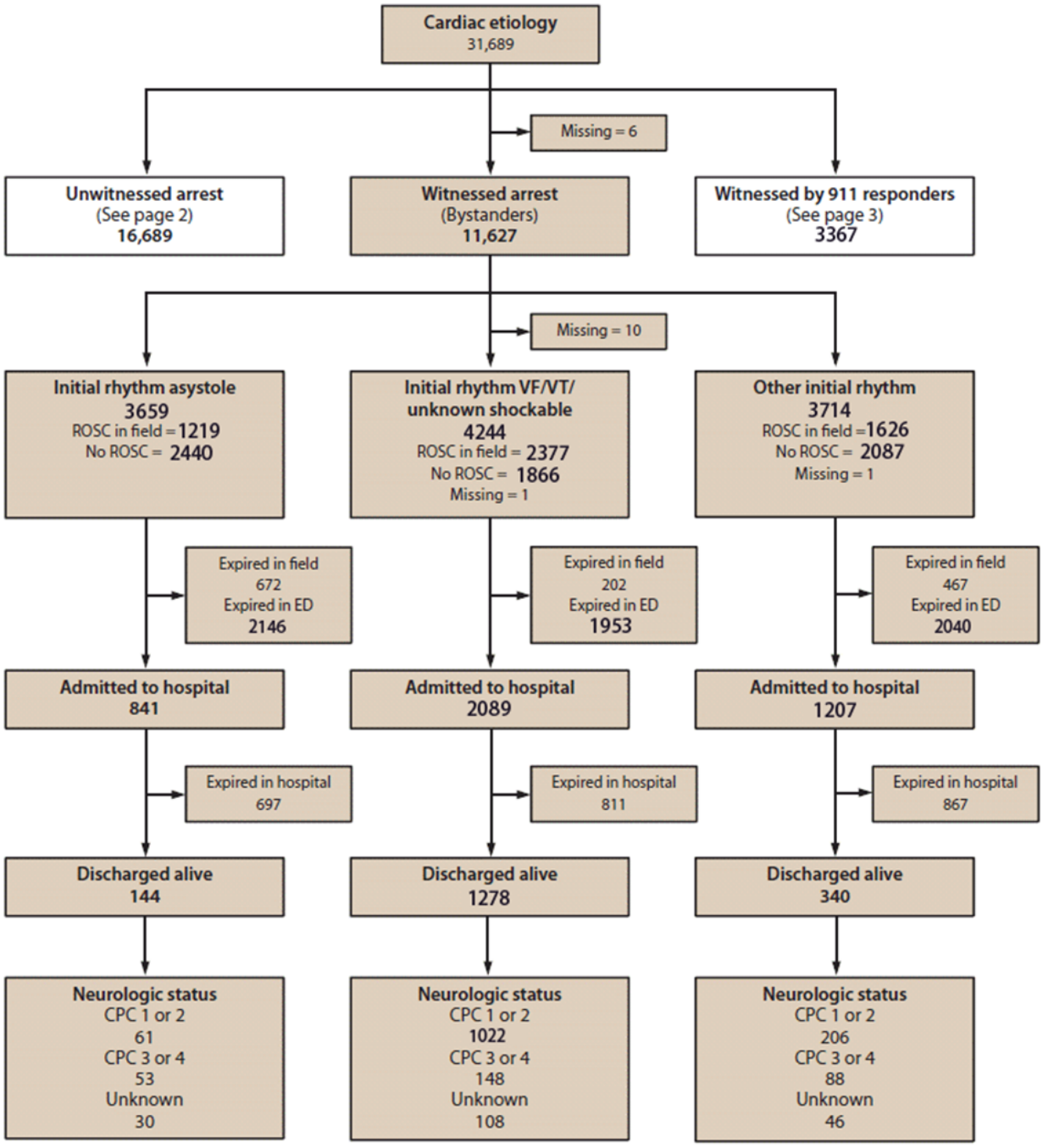


Network Chain of Complex Systems Centered on Cardiac Emergency Decision Making

Definition of a Complex System Network Chain

The information chain of heart disease emergency treatment can be described as an information set in the information source that conforms to certain time-dependent data constituting an information transmission chain structure and the information chain of emergency decision making during a heart attack (*link i,i = 1,2,…,n*) (Figure 2) [4]. The emergency treatment process of the heart disease itself is a feedback system composed of multitasking and time nodes. In this system, each data point forms a dynamic fusion, cooperation, and feedback process of multiple devices and information sources.

Figure 2. Network structure of a complex system centered on cardiac emergency decision making. VF: ventricular fibrillation, VT: ventricular, ROSC: return of spontaneous circulation, ED: emergency department; CPC: cerebral performance category.



Network Chain Composition of a Complex System

The complex system network chain structure centered on heart disease emergency decision making is mainly composed of an emergency information resource system, a first aid rule system, a cardiac emergency rescue system, a medical staff organization structure, a monitoring system (which consists of six subsystems like aldosterone), and a diagnostic system. For example, the organizational structure of medical staff includes attending physicians, associate attending physicians, perfusionists, clinical engineers, etc, and information is connected and shared between the various subjects [5].

The entire complex system network chain structure is composed of three interactive information chains. Information chain #1 includes an emergency information resource system, clinical path, and a diagnostic system; information chain #2 includes an emergency information resource system, clinical path, and a diagnostic system; and information chain #3 includes a cardiac emergency care system, medical staff organizational structure, and a diagnostic system. Each information chain has a separate subnet chain, and through the transmission channel Ch_1, the complex system network chain structure can be shared and called. For example, the vital sign parameters of the patient are obtained from the emergency decision information resource system, including the attributes of the patient's vital signs parameters such as temperature, blood pressure, heart beat (f_1,f_2,…,f_n), etc, to provide emergency decision information for the cardiac emergency specialist. Regarding the information chain of emergency decision making in cardiac emergency, due to the restriction of the resistance coefficient of channel network Ch_1, the information transmission process needs to pass as few channels as possible to achieve the most efficient and seamless transmission of information.

Definition of State Space for Cardiac Emergency Decisions

The entire first aid process is a large complex system. The various data, information, and knowledge generated and used in it are employed as resources to support first aid decision making. They constitute a multilayer, heterogeneous, and massive complex information space. To represent this state, state space equations can be used, and subspaces can be transformed. A state space equation composed of matrices is a widely used form in subspace identification. Most of the subspace identification algorithms are carried out using subspaces and their mapping methods. In this paper, orthogonal projection is used to obtain the vector space for the state of first aid for heart disease [6].

The static structure (L) of the cardiac emergency decision-making system consists of the emergency information space M, the emergency reasoning space Q, and the emergency decision space D. The cardiac emergency information space is a vector space defined on a real or complex field F as M. It consists of dimension space, and dimension space contains collection space, and collection space contains attribute space. Therefore, M can be represented by the following quaternion:

$$M(x) \le Z, Se, At, Qij \quad (1)$$

Among them, Z, Se, and At represent dimension space, set space, and attribute space, respectively, and are composed of the emergency medical information vector. The dimension space constituting the emergency response decision M mainly includes the case dimension space Z_c of case-based reasoning (CBR), the rule dimension space Z_r of rule-based reasoning (RBR), the resource dimension space Z_e of decision reasoning, and the time dimension space Z_t of decision reasoning. The dimensional space is multidimensional, and the contained collection space or attribute space is heterogeneous. For example, the information in the vital sign collection Se_v in Z_c is different from the automatic control information in the medical emergency equipment collection Se_{QU} in Z_e. We introduce the definition of emergency reasoning space given in the cardiac emergency decision system below.

The knowledge structure composed of all the data, information, and knowledge of the subdecision process in the cardiac emergency information space is called the emergency reasoning space, which is represented by Q. Q can be expressed as a quaternion:

$$Q = (U, A, V, C) \quad (2)$$

Among them, U represents the fusion first aid reasoning space metadata, A is the feature set of the first aid reasoning space; its feature value range is V,  , where u_i is the feature value vector, and C is the solution of the first aid reasoning space. In Q, an inference information function is formed between C and U:

$$f: U \times A \rightarrow C \quad (3)$$

According to the properties of the corresponding space vector in the state space, the above relationship shows that Q can be divided into several subspaces. For Q, according to the definition of the state space for cardiac emergency, there are:

There is an invertible matrix S_i,$I = 1,2,\ldots,n$ such that $A - BD^{-1}C = S_iA^{-T}S_i^{-1}$.

Cardiac Emergency Decision Data Space Framework

In space three, the information space for cardiac emergency decision making is constructed. The dimension is an item-level division of the entire cardiac emergency decision-making system, including case dimension (CBR), rule dimension (RBR), resource dimension, and time window. The concept of subsets is divided under the dimensions, and there is a corresponding subset under each dimension. The next level of attributes describes these subsets and their relationships in detail. Different subsets contain different attributes, and the subset attributes of different dimensions are described in different ways. This vector space system reflects the most important characteristics of cardiac rescue and is the core foundation for the establishment of a cardiac rescue model [7]. The ternary composition of M is a "dimension-subset-attribute" frame (digital subtraction angiography frame), which describes the entire dimension space of the first aid of the heart by constructing a dimensional matrix space. The concept of subset space is used to describe the content of each dimension space, and the attribute space is used to construct a local model frame in each subset space.

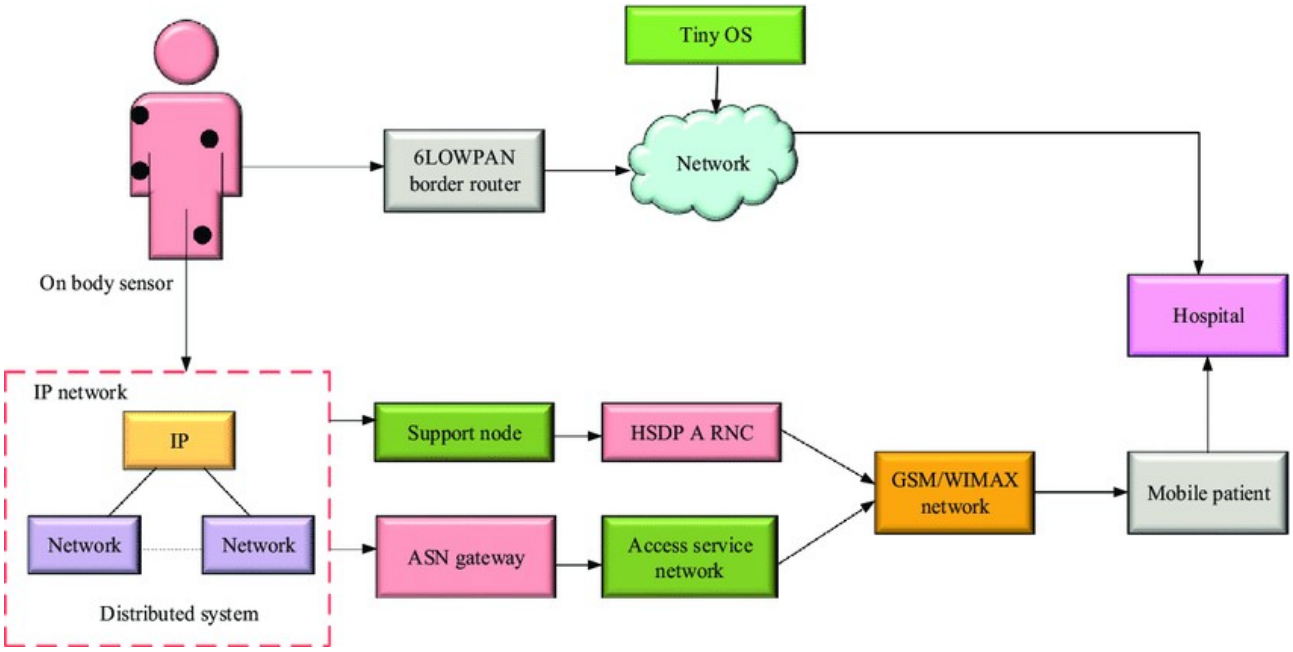
Dimension space (Z) is an 4D solid vector space structure model used to describe cardiac emergency information space. It mainly includes case-dimensional space, regular-dimensional space, resource-dimensional space and time-dimensional space. Z can be expressed as:

Among them, Z_c, Z_r, Z_e, and Z_t refer to the case dimension, rule dimension, resource dimension, and time dimension, respectively.

Subset space (*Se*) is a vector space structure model about a subset of features that exists in any dimension space in the 4D cardiac emergency information space. It can be seen that Z_c is a dimension space formed by and represents the richness of the emergency decision-making experience of the first-aid heart disease specialists. Attribute space (*At*) refers to the specific combination of attribute vectors that constitutes the space of each subset, and can be expressed as:

In the formula, $at_{di}(se)$ represents the attribute component, subscript di represents the dimension, and se represents the subset component. Take the physical collection space Sev as an example to illustrate the correlation. Let body temperature, blood pressure, heart beat, and other attributes be denoted as $a_1,a_2,a_3,\ldots,at_{\mathrm{CBR}}$ (Se_v), the attribute vector of the knife dimension, so at_{CBR} $(Se_v) = \{a_1,a_2,a_3,\ldots,a_n\}$. The Dimension One Subset One Attribute framework of the cardiology emergency information space is shown in Figure 3. In Figure 3, a knowledge representation model of the state space for cardiac emergency response is constructed using the "dimensional-subset-attribute" framework [8].

Figure 3. The "dimension-subset-attribute" framework of the cardiac emergency information space. OS: overall survival, IP: intraperitoneal, ASN: asparagine, GSM: Global System for Mobile Communications.

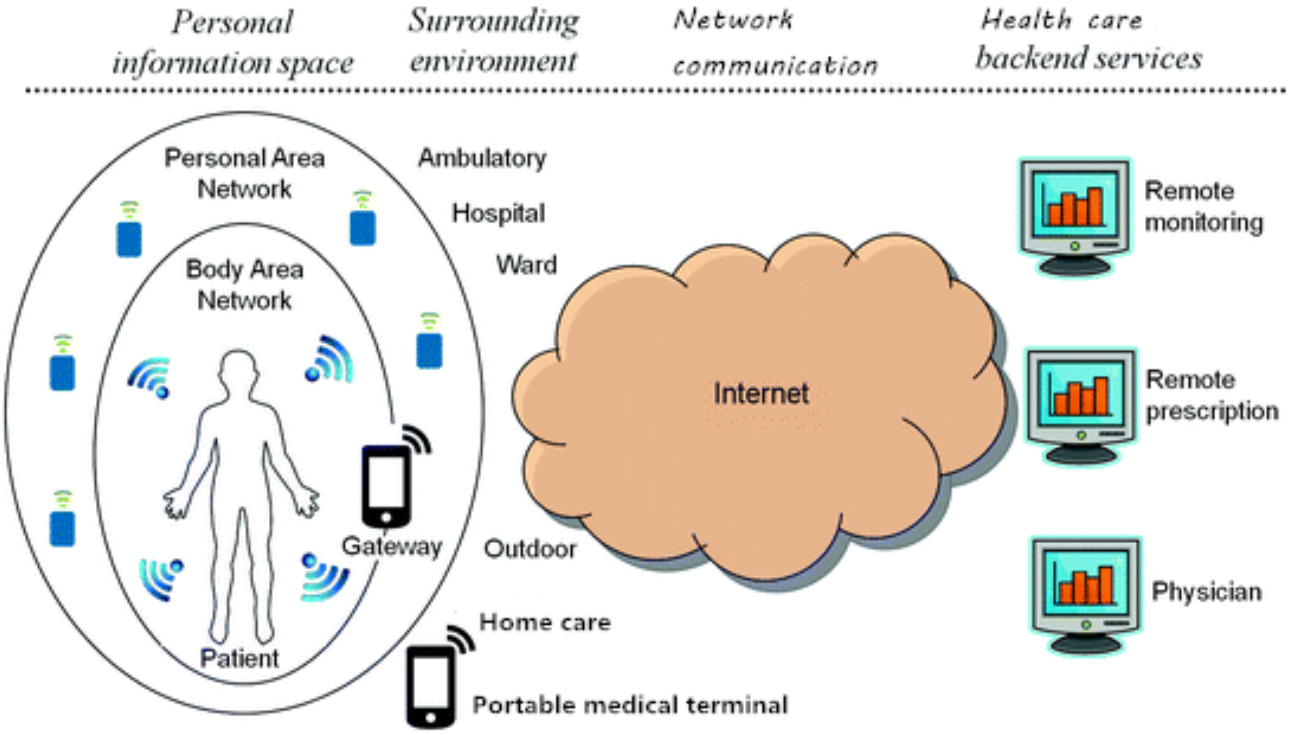


Dynamic Performance of the Cardiology Emergency Reasoning Space

The process of acquiring and transferring knowledge in the first aid reasoning space is complicated. In order to study the robustness of the first aid decision-making system, the dynamic performance of the first aid reasoning space needs to be studied. From these spaces containing decision-making reasoning information, a knowledge space (knowledge base) for decision-making processes is collected, extracted, and mapped from the cardiac emergency information subspace to the reasoning knowledge space. In the multidimensional cardiac emergency information space, the data sets required for decision-making reasoning are obtained and transmitted, and these data constitute heterogeneous knowledge of decision-making reasoning.

The neutron space of complex cardiac emergency decision making is recessive, random, or uncontrollable, and the information generated shows system characteristics such as multidimensional, dynamic, and uncertain. In terms of inference knowledge transfer, for all the attributes obtained in the decision-making process of cardiac emergency, using attribute reduction, the initial inference state space *S* formed, eliminating attribute redundancy and retaining the relevant attribute set of information, reasoning for t temporal. The space *S* is used for the information characterization of relationship mapping, the role of reasoning, or the influence of uncertain factors to obtain the $t + \Delta t$ temporal reasoning state S^n. In this way, the N_s step reasoning is performed, and finally, the emergency decision-making knowledge *C* is obtained, as shown in Figure 4 [9].

Figure 4. The state chain of the cardiac emergency reasoning space.



Regarding the state transition chain, for the two information state subspaces $S = (U, AT, V, f)$ and $S' = (U', AT', V', f')$ of the state at any t and $t + \Delta t$ in the emergency reasoning space Q, where $AT, AT' \subset A$, S and S' are the two reasoning state subspaces in Q.

Methods

Robustness of Cardiac Emergency Decisions

Definition of Robustness

System robustness is a characteristic that maintains some of its original performance under certain internal or external parameter perturbations. The robustness complements the system's vulnerability to ensure the overall safety of the system, which can reduce the uncertainty impact caused by the errors or parameter errors of the first aid reasoning model of cardiac disease. System fragility and system stability and robustness are two aspects of the same problem. Things that are robust in one aspect may be vulnerable in another aspect, or be robust in one set and vulnerable in another.

From an application perspective, the objects of robust control have been extended from the initial basic control problems to multiple areas such as optimal control, system identification, decision making, and reasoning. Many scholars have carried out in-depth research work in the theoretical and application stages of robust control, which has extended the robustness research to many specific application levels, such as network theory, ecology, computer theory, mechanical vibration theory, pattern recognition, and reasoning. In the fields of prediction and system modelling, robustness research has achieved different degrees of results and has been applied in physical systems. Most indicators that measure a system comprehensively can be attributed to the system's stability (anti-interference ability), accuracy (steady-state error), and fastness (response speed). Cardiac emergency decision-making robustness can be divided into static performance and dynamic performance according to its state. Static performance refers to the state characteristics of the system when various factors inside the system, system input, and external interference are constant. Specifically, it includes system stability, steady-state error, and anti-interference. Dynamic performance refers to when the system has various internal factors. When factors, system inputs, or external disturbances change, the characteristics of the transition process of the system from one state to another, including the speed of response. These static and dynamic properties are interrelated and affect each other, and together they constitute the robustness of cardiac emergency decision making.

Uncertainty in Cardiac Emergency Decisions

Under the condition of information uncertainty or system disturbance, emergency decision making in cases of heart disease shows vulnerability and lacks robustness. Certainty means that a cardiac emergency specialist has sufficient information of adequate quality to accurately describe, reason, or make emergency decisions about the behavior or other characteristics of heart disease in a quantitative manner. What cannot be described by the above definition is called uncertainty. Cardiac emergency decision-making uncertainty refers to the hidden, random, or uncontrollable characteristics of the subspaces involved in emergency response in the decision of cardiac emergency, such as the difference in the medical level of medical personnel, technical limitations, and medical emergency equipment data errors, etc. The uncertainty of cardiac emergency decision making is reflected in Table 1. Turning these uncertain situations into deterministic situations requires more or better information.

Table 1. Uncertainty in cardiac emergency decision making.

Performance of uncertainty	Features	Examples
Case dimension	Recessive, random	Patient system uncertainty, patient's own physical conditions, and existing medical technology limitations
Regular dimension	Random, uncontrollable	Uncertainty of multiple participants in the operation, uncontrollable emergencies in the operating room, and multiple personnel participating in the operation
Resource dimension space	Hidden and uncontrollable	Uncertainty of medical emergency equipment, raw errors of instrumental measurement numbers, errors caused by mutual electromagnetic interference between multiple instruments
Time dimension space	Dynamic and uncontrollable	The time constraints faced by emergencies and the uncertainty caused by system interference, such as the dynamic process of optimal schemes

Robustness Constraints

Robust stability criteria and sensitivity functions are the basis for further research on the robustness of cardiac emergency decision making. To this end, it is necessary to determine the robust constraints of cardiac emergency reasoning. Let v be the value of all deterministic variables in the model, such as the age attribute value of heart disease cases, etc; $M(v)$ is a robust random mixed model; $M_0(v)$ is a function matrix determined solely by deterministic variables, and $M_1(v)$ and $M_2(v)$ are uncertainty weight function matrix for random terms; θ is a random term for uncertainty; and I is an identity matrix.

Results

Modeling of CBR / RBR Fusion Inference Based on Robustness Threshold

Robust Solution Set for the Fusion Inference Space

The complex relationships in the knowledge source are mainly reflected in the three perspectives of similarity discrimination relationship, rule discrimination relationship, and inference reliability. The singular value of the fusion space is an intrinsic characteristic quantity calibration in the decision-making process of cardiac emergency, and a 3D robust threshold vector is constructed with the singular value.

The ternary vector constructed by the case attribute value matrix P, the regular membership matrix Q, and the normalized singular value is called the inference robustness threshold vector.

The singular value is used as the robustness threshold, and the inherent characteristics of the quantity matrix from case knowledge and rule knowledge are considered, which not only represents the correlation between the metadata of the knowledge source and the solution set of the decision target, but also reflects the threshold in the fusion space The internal differences overcome the strong subjectivity in the previous fusion methods, have strong robustness, and effectively improve the reliability of reasoning.

Fusion Reasoning Strategies and Steps

Considering the similar accumulation and reuse of knowledge sources, the simple CBR consists of main five steps, including the following: problem representation, case retrieval, solution transfer, feature mapping, and adjustment of noncorresponding solutions, and problem identification with RBR, generation of proposed solutions, recommendation evaluation, and screening. Based on a four-step method, including scheme modification, a robust threshold CBR is proposed. For the RBR fusion inference method, the specific implementation steps are as follows:

- Step 1: Knowledge representation and problem identification. A case base CB and a rule base RB are generated on the knowledge source (N,Q) of the unitary space E. The target problem is identified, and the target problem set Cq is established, and the case matrix R, the rule matrix S, the robust solution x* of fusion inference, and the inference operator matrix T(x) are left blank [10].
- Step 2: Fusion spatial mapping and normalization processing. Normalize the case data and rule knowledge Vij in the knowledge source (N,Q). The orthogonal discriminant method is used to obtain R for the case set, and the decision tree rules and S are obtained to form the fusion unitary space E of fusion reasoning.

- Step 3: Determine robustness threshold and fusion inference strategy. Using singular value decomposition, the robustness threshold vector is obtained, and the relationship between the target problem Cq and the meta-knowledge in the knowledge base is defined in E, and then the fusion inference strategy is formulated.
- Step 4: Integrate reasoning in knowledge space E and iterate repeatedly until a robust solution is obtained.

Experimental Research and Application Effect Analysis

The application effect of this thesis includes both theory and application. In terms of theory, based on mainstream research such as magnetic induction tomography and this hotspot of medical intelligent reasoning, a set of mathematical models such as the R2MIFS (represents the treatment mode of fibroblastic sarcoma) model and the RTCRF (clinical trial observation mode) model are established to solve the problem of robustness of cardiac emergency decision making. The R2MIFS model can effectively eliminate the attribute feature redundancy of patient characteristic data; the RTCI (the antimicrobial peptide Rana temporaria chensinensis) model can effectively implement resource strategy separation and conflict resolution mechanisms, implement inference knowledge fusion,

and have obvious advantages in efficiency and accuracy compared with foreign literature; the BN (Bayesian network)-CBR / RBR model enhances the accuracy and sensitivity of inference decision making. In clinical trials, it has been experimentally verified that it has a good decision-supporting effect on the diagnosis of heart disease.

Discussion

From the perspective of the robustness of complex medical decision-making systems, our paper summarizes the evolution of related research content in the field over the years and the research results obtained, and proposes a three-level perspective on the research process. They are as follows: (1) research at the "behavioral level" based on industrial engineering theory (ie, medical industrial engineering); (2) research on the robustness of complex systems based on cybernetics (ie, medical system engineering); (3) "logical level" research based on information theory and artificial intelligence (ie, medical artificial intelligence).

Acknowledgments

This project was supported partially by the Population Health Informatization in Hebei Province Engineering Technology Research Center, Hebei North University (#JYT2019016) and the Zhangjiakou Municipal Science and Technology and Seismological Bureau (#1911016C-9).

Conflicts of Interest

None declared.

References

1. World Health Organization. 2004. World Alliance For Patient Safety Forward Programme 2005 URL: https://www.who.int/patientsafety/en/brochure_final.pdf [accessed 2020-06-09]
2. Jeganathan J, Knio Z, Amador Y, Hai T, Khamooshian A, Matyal R, et al. Artificial intelligence in mitral valve analysis. Ann Card Anaesth 2017;20(2):129-134 [FREE Full text] [doi: 10.4103/aca.ACA_243_16] [Medline: 28393769]
3. Vancini-Campanharo CR, Vancini RL, de Lira CAB, Andrade MDS, Lopes MCBT, Okuno MFP, et al. Characterization of cardiac arrest in the emergency department of a Brazilian University Reference Hospital: A prospective study. Indian J Med Res 2016 Oct;144(4):552-559 [FREE Full text] [doi: 10.4103/0971-5916.200898] [Medline: 28256463]
4. Bronzi W, Frank R, Castignani G, Engel T. Bluetooth Low Energy performance and robustness analysis for Inter-Vehicular Communications. Ad Hoc Networks 2016 Feb;37(P1):76-86. [doi: 10.1016/j.adhoc.2015.08.007]
5. Li Y, Wang X, Li H, Wang J, Zhu X, Li Y, et al. SU-F-T-192: Study of Robustness Analysis Method of Multiple Field Optimized IMPT Plans for Head & Neck Patients. Med. Phys 2016 Jun 07;43(6Part15):3506-3506. [doi: 10.1118/1.4956329]
6. Lee H, Troschel FM, Tajmir S, Fuchs G, Mario J, Fintelmann FJ, et al. Pixel-Level Deep Segmentation: Artificial Intelligence Quantifies Muscle on Computed Tomography for Body Morphometric Analysis. J Digit Imaging 2017 Aug;30(4):487-498 [FREE Full text] [doi: 10.1007/s10278-017-9988-z] [Medline: 28653123]
7. Kharin AY. An Approach to Asymptotic Robustness Analysis of Sequential Tests for Composite Parametric Hypotheses. J Math Sci 2017 Oct 11;227(2):196-203. [doi: 10.1007/s10958-017-3585-z]
8. Iglesias López S, Llopis García G, Yañez-Palma MC, Rodríguez Adrada E. [Detection of palliative patients with acute cardiac insufficiency in the emergency department]. An Sist Sanit Navar 2016;39(2):323-324 [FREE Full text] [doi: 10.23938/ASSN.0259] [Medline: 27599962]
9. Bhoi S, Mishra PR, Soni KD, Baitha U, Sinha TP. Epidemiology of traumatic cardiac arrest in patients presenting to emergency department at a level 1 trauma center. Indian J Crit Care Med 2016 Aug;20(8):469-472 [FREE Full text] [doi: 10.4103/0972-5229.188198] [Medline: 27630459]
10. Aguiar PR, Da Silva RB, Gerônimo TM, Franchin MN, Bianchi EC. Estimating high precision hole diameters of aerospace alloys using artificial intelligence systems: a comparative analysis of different techniques. J Braz. Soc. Mech. Sci. Eng 2016 Mar 19;39(1):127-153. [doi: 10.1007/s40430-016-0525-7]

Abbreviations

BN: Bayesian network
CBR: case-based reasoning
CPR: cardiopulmonary resuscitation
RBR: rule-based reasoning
RTCI: Rana temporaria chensinensis
VF: ventricular fibrillation
VT: ventricular

Edited by K Kalemaki, Z Du, H Li; submitted 16.04.20; peer-reviewed by J You, S Yao, H Zhang; comments to author 30.04.20; revised version received 03.05.20; accepted 11.05.20; published 27.07.20.

Please cite as:
Gong L, Zhang X, Li L
An Artificial Intelligence Fusion Model for Cardiac Emergency Decision Making: Application and Robustness Analysis
JMIR Med Inform 2020;8(7):e19428
URL: https://medinform.jmir.org/2020/7/e19428
doi:10.2196/19428
PMID:32716305

Viewpoint

The Role of Health Technology and Informatics in a Global Public Health Emergency: Practices and Implications From the COVID-19 Pandemic

Jiancheng Ye[1]

Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

Corresponding Author:
Jiancheng Ye
Feinberg School of Medicine
Northwestern University
633 N Saint Clair St
Chicago, IL
United States
Phone: 1 312 503 3690
Email: jiancheng.ye@u.northwestern.edu

Abstract

At present, the coronavirus disease (COVID-19) is spreading around the world. It is a critical and important task to take thorough efforts to prevent and control the pandemic. Compared with severe acute respiratory syndrome and Middle East Respiratory Syndrome, COVID-19 spreads more rapidly owing to increased globalization, a longer incubation period, and unobvious symptoms. As the coronavirus has the characteristics of strong transmission and weak lethality, and since the large-scale increase of infected people may overwhelm health care systems, efforts are needed to treat critical patients, track and manage the health status of residents, and isolate suspected patients. The application of emerging health technologies and digital practices in health care, such as artificial intelligence, telemedicine or telehealth, mobile health, big data, 5G, and the Internet of Things, have become powerful “weapons” to fight against the pandemic and provide strong support in pandemic prevention and control. Applications and evaluations of all of these technologies, practices, and health delivery services are highlighted in this study.

(JMIR Med Inform 2020;8(7):e19866) doi:10.2196/19866

KEYWORDS

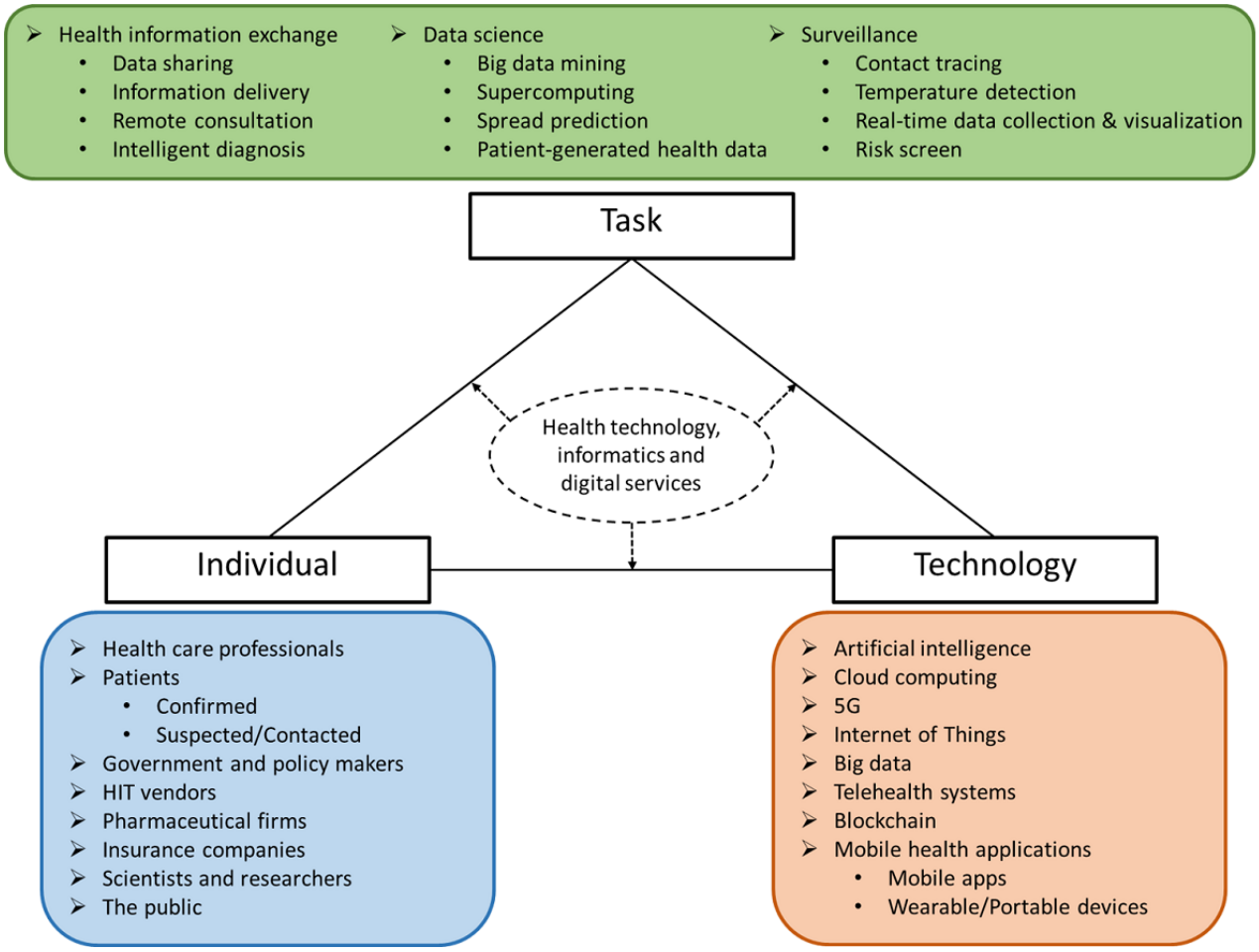
health technology; health information system; COVID-19; artificial intelligence; telemedicine; big data; privacy

Introduction

In December 2019, an emerging infectious outbreak was found in Wuhan, Hubei Province, China, and it was caused by the novel coronavirus (2019-nCoV) [1,2]. At present, the pandemic is spreading across the whole world [3]. There has been human-to-human and health care worker transmission, but the source of the coronavirus disease (COVID-19) has not been found; the route of pandemic transmission has not been fully understood. The virus may mutate, and further spread is almost certain. 2019-nCoV has a long incubation period and strong infectivity; therefore, the prevention and control of the COVID-19 pandemic faces great challenges. Compared with severe acute respiratory syndrome (SARS) and Middle East Respiratory Syndrome (MERS), COVID-19 has some new and different features; it has spread more rapidly due to increased globalization, a longer incubation period, and hidden symptoms [4].

Integrating novel health technologies and practices such as artificial intelligence (AI), big data, 5G mobile networks, Internet of Things (IoT), mobil health applications, telehealth services, and health information exchange (HIE) services [5] into health care systems can aid in the following ways: reporting and monitoring human transmission information, data assortment and analysis, tracking and sending alarms, etc. All of these functions are helpful to provide strong support in pandemic prevention and control. Figure 1 demonstrates a framework of health technologies, informatics, and digital services following a modified fit between individuals, task, and technology (FITT) framework [6].

Figure 1. Modified fit between individuals, task, and technology (FITT) framework for health technologies, informatics, and digital services. HIT: health information technology.



Roles and Capabilities of Health Technology and Informatics in the COVID-19 Pandemic

Taking advantage of health technologies in assisting the investigation and judgment of the pandemic will innovate diagnosis and treatment, improve service efficiency, and strengthen the capacity of information technology to support pandemic prevention and control. The specific schemes are described in the following sections.

Health Information Acquisition and Data Analysis Application

To carry out the real-time development of pandemic tracking, key screening, and effective prediction, we should make full use of disease prevention and control information systems, and actively adopt the method of online pandemic information reporting. Big data technology will provide support for scientific prevention and policy development. The health care systems need to form a multisource platform that integrates data monitoring, exchange, convergence, and feedback mechanisms for road, railway, civil aviation, communications, medical, and other pandemic-related parties. The integrated platform will enhance the information linkage with public security, transportation, and other departments. COVID-19 diagnosis and suspected medical record collection need to be used to conduct analysis and application from the regional and national health information platform, which aids in the pandemic prevention and control, clinical treatment, and scientific research.

Telehealth Services

With the assistance of health information technology (HIT), major hospitals, along with designated hospitals, can provide services such as remote consultation and prevention and control guidance. This will improve the ability of elementary health institutions to deal with the pandemic situation and relieve the pressure of designated hospitals.

Moreover, internet hospitals or internet diagnosis and treatment platforms have their advantages. They can provide online rediagnosis of some common and chronic diseases and drug distribution services, and reduce the risk of cross infection of patients' offline visits. Public and standardized platforms should be adopted to gather the service links of registered and approved internet hospitals and internet diagnosis and treatment websites. This adoption would make it more convenient for people to obtain timely information on pandemic prevention and control, and diagnosis and treatment services. Organizing health care institutions at all levels to provide online compulsory counseling, home medical observation, and guidance for pneumonitis infected by COVID-19 will expand the online medical service

space, guide patients to seek medical treatment, and relieve the pressure of offline clinics.

Scientific Popularization and Real-Time Information Disclosure

COVID-19 knowledge popularization and prevention of transmission is critical to facilitate timely access to authoritative information, and assist in understanding the diseases and self-protection scientifically. Adopting national integrated online service platforms and official websites of health administration departments at all levels will be helpful to carry out pandemic information inquiries.

Social media and other popular platforms should be employed to carry out training on self-protection and for providing COVID-19 diagnosis and treatment education. This will improve the health care services and personal protection capabilities of primary care institutions.

Health Information System Deployment

To ensure the smooth operation of the pandemic prevention and control system, an upgrade and transformation of the network should be accelerated. Where appropriate, information technology such as 5G should be applied to improve the stability and transmission quality of the designated hospitals' network to meet the needs of patient treatment.

On the other hand, pandemic monitoring and analysis, virus tracing, patient tracking, personnel flow, and community management are important tasks during the pandemic. Health technologies such as big data, cloud computing, and AI can be applied to develop scientific strategies for precision prevention and control of the pandemic. Internet platforms can help to match the supply and demand of medical and pandemic prevention materials precisely. In this way, coordinated deployment and recycling management can be achieved. Information technology enterprises and medical research institutions should work jointly to tackle key problems, accelerate the detection and diagnosis of COVID-19 infection, and conduct research and development of new vaccines.

Health Technology Applications and Digital Service Practices

Artificial intelligence (AI)

As 2019-nCoV has the characteristics of strong transmission and weak lethality, the large-scale increase of infected people may drag down the medical system, so the pandemic prevention and control needs to track and manage the health status of residents and isolate suspected patients. AI's capacity of in-depth mining and processing massive information has been used to detect and predict the spread of viruses in pandemic situations [7,8] by building intelligence monitoring platforms and developing advanced algorithms such as neural networks, the susceptible-exposed-infectious-removed (SEIR) model, and long short-term memory (LSTM) networks [9-12]. For instance, data from social media, contact tracing, surveys, etc [13] could be applied to various machine learning or deep learning models to predict the course of COVID-19 and potential reappearances [11].

After the outbreak of COVID-19, the false negative outcome of the kit test increased the difficulty of diagnosis. AI has become a powerful supplement to kit detection. The increasingly mature AI medical imaging technology [14], through tagging a large number of medical image samples and applying them to the algorithms for training, learning, and understanding, could effectively assist doctors in decision making. Furthermore, AI is also on the frontlines of the pandemic. For instance, intelligent robots are collections of integrated multi-sensor fusion, path planning, robot vision, intelligent control, and human-computer interface technology. They can provide diverse services such as disinfection, food delivery, and medicine delivery. In the situation of scarce protective clothing, the pressure on the frontline health care workers to diagnose and treat can be relieved to a certain extent, and the chance of cross-infection can be reduced.

Telemedicine, Telehealth, and Telecommunications Technology

Telecommuting has been widely used in industries and businesses by replacing traditional commuting with digital technologies [15,16]. In the health care domain, telemedicine has been leveraging telecommunications technologies to make use of HIT and medical information such as video imaging, thus, allowing health care providers to work remotely [17,18]. Telehealth has been interchangeably used with telemedicine [19,20], but as an umbrella term, telehealth incorporates the functions of telemedicine and a variety of nonclinical services like tele-pharmacy and tele-nursing [21]. The emergence of telehealth in the coronavirus pandemic dates back to SARS in 2003 [22]. During the SARS period, to reduce people's mobility and cross infection, a lot of work was conducted through the internet, telehealth began to show its usefulness. However, due to the network quality and technical level at that time, telehealth was limited to simple online or telephone consultation, which was relatively elementary.

Relying on the rapid development of the telecommunication technologies, video imaging, 5G, and other technologies in recent years, telehealth has developed rapidly. At present, the scenarios of remote expert consultation and remote medical education have been widely applied. Telehealth can solve the problem of unbalanced development of medical resources among regions. Through the internet, experts can organize remote consultation in remote areas and resource-deficient areas. Textbox 1 presents two basic models of telehealth.

Telehealth has brought the benefits of high-quality medical resources from superior hospitals to the community-level health institutions, and it has greatly improved health quality, efficacy, efficiency, and saved expenses.

Remote diagnosis through internet technology and big data enables patients to perform imaging and electrocardiogram examinations at township-level hospitals, and doctors at grassroots hospitals can transmit information to their superiors via the internet. The regional diagnosis center and experts will issue a timely diagnosis report, which is convenient for doctors at the elementary hospital to provide patients with targeted treatment.

During the COVID-19 pandemic, telehealth systems can assist and support health care professionals to conduct remote consultation so that doctors from different regions can converge together to discuss the diagnosis and treatment of patients [23]. The system will connect the remote mobile workstation beside the patient's bed, and real-time video meetings can be carried out so that the experts can observe the actual situation of the patient. During the process, patients have interactions through video [24] or telephone [25] with health care providers [26]. The history of symptoms and exposure risk will be obtained through an observational assessment. Based on this information, doctors can make judgments on whether the patient has been infected or needs further testing. Telehealth has also been used to help manage the patients with suspected symptoms of COVID-19 and provide virtual medical services for chronic diseases [27,28]. The image diagnosis system digitizes the results generated by the examination equipment (x-ray, ultrasound machine, etc) in the medical institution; health care providers can access the electronic health records (EHR) remotely through the network to realize remote diagnosis. For patients who are critically ill, telehealth intensive care equipment transmits the physiological information and medical parameters to the monitoring center through the telecommunication network, and real-time detection and further analysis can be conducted. The systems can realize interoperability of automated health data through HIE [5] and the sharing of medical information with participating hospitals. Telehealth shortens the distance between doctors and patients, and helps doctors provide timely medical services based on physiological information transmitted from distant places [29]. Experts from other places remotely access the EHR of patients in the insulation ward, discuss the treatment plan and effect in real time, and give professional opinions.

Given the shortage of medical protective material and personal protective equipment (PPE), the use of remote consultation reduces the occurrence of on-site diagnoses, which also saves PPE, and reduces the risk of infection spread caused by the transfer of diagnosed patients to the superior hospitals.

Textbox 1. Basic models of telehealth.

Direct docking mode

If the inviting medical institution finds the invitee on its own, then telehealth can be carried out directly between the medical institutions.

Platform matching mode

If the inviting medical institution cannot find the invitee by itself, it can publish the requirements on the remote medical service platform established by the inviting medical institution or a third-party institution. The platform can match the invitee's or other medical institutions' initiative to the inviter's needs of health response.

5G

The fifth generation mobile networks, or fifth generation wireless systems, is the latest generation of cellular mobile communication technology. It provides at least 10 Gbit/s peak rate and millisecond-level transmission delay [30]. The network enables a new kind of network that is designed to connect everything together virtually, including human and machines, objects and devices with high reliability and capacity. Compared to prior generations, 5G has larger bandwidth, higher rate, lower delay, and larger connection. 5G communication technology can not only achieve high-quality transmission of 3D images but also provide services such as data acquisition, real-time positioning, remote diagnosis and treatment, and other fusion functions in addition to information communication [31].

In a prior generation of wireless systems such as 4G, network bandwidth was limited, only meeting the transmission of small volumes of medical information. For medical images like computed tomography scan images [23], real-time remote consultation, and telemedicine meetings, the network must have high transmission speed and low latency. The COVID-19 pandemic is a global emergency; saving time means saving patients' lives.

In the medical industry, communication is one of the important factors that impact the development of medical rescue. The high-speed communication capacity of 5G can effectively improve the efficiency of medical emergency rescue and the response ability at public health events.

Remote diagnosis and treatment based on 5G between doctors and patients can be realized through real-time high-definition audio and video connection. Medical data transmission can promote the transition from "face-to-face" consultation to video remote consultation, which further improves efficiency and precision. Under the current situation of the pandemic, 5G and telehealth can make diagnosis and treatment more efficient, convenient, and safe. The high-speed, large-capacity, and low-latency of the 5G network accelerate achieving the needs of real-time, high efficiency, and stability of the remote consultation.

5G telecommunication technologies also improve the health care accessibility. Health information systems gather clinical data from various locations such as hospitals, community health care organizations, and physician practices. 5G supports real-time health data exchanging so that health care professionals can get access to patients' diagnosis records, medical history, and lab results without delay and information transmission barriers. Prior generation of wireless systems could not provide such capacities for HIE [32] and may cause health care workforce burnout [33]. With the implementation of 5G, the traditional medical workflow will be improved dramatically, and the unnecessary contact between health care providers and patients may be reduced. The electronic health information system can track the whole process of the medical order, which decreases the risks of medical errors and improves the quality of health care and system management.

In the prevention and control of COVID-19, it is essential to make full use of 5G communication technology, linking doctors

and experts across the country and even around the world; actively taking advantage of online diagnosis and treatment; carrying out online consultation, health science popularization, psychological assistance counseling, and home isolation guidance services such as delivery of medicines for chronic diseases; and enabling patients to receive health care without leaving home, all of which will reduce the risk of being exposed to the virus.

Internet of Things (IoT)

IoT is a novel paradigm of interrelated digital machines, mechanicals, computing devices, and other objects [34,35]. Based on the communication protocols, it combines with the internet to realize the intelligent management of information. By taking advantage of communication technologies such as networks or the internet and sensors, everything could be linked together to realize the connection between people and objects or objects and objects. At the same time, people-oriented information, remote monitoring and control, and intelligent management are realized. Remote monitoring [36] based on IoT is an effective way to assist in achieving real-time, continuous, and long-term monitoring of patient vital signs and transfers the acquired health data or critical alarm information to health care professionals. Remote monitoring can also achieve real-time data acquisition and analysis of patients in isolated areas. Heart rate, breathing, and other physiological indicators [37] of patients with COVID-19 can be collected in the isolated zone through smart devices [38], electrocardiographs, ventilators, and sphygmomanometers; data can then be sent back in real time through Bluetooth [37] or the network. The collected data of patients' physiological signs can be analyzed and processed intelligently in the clinical decision support system. Once the system finds abnormal data, it will send an alarm [39] in real time, and doctors will investigate and judge the situation according to the alarm information.

Mobile Health Apps

Patient-generated health data (PGHD) are becoming more and more important in health care, especially in the COVID-19 pandemic. Coronavirus tracking apps on smartphones are playing critical roles as a mobile technology to collect, gather, and share PGHD during the pandemic. Users can check whether they have had contact with patients who are infected by entering personal information. When registering, users are required to enter their name as well as age, zip code, and other relevant information. Phone numbers are also recorded, so users will be notified once the system identifies contact with patients who are infected. Most trace tracking apps use network or Bluetooth technologies. When other users get close to a certain social distance range, the app will perform digital handshakes with them; this "interaction" will be recorded and encrypted [40]. This information is useful to identify high- and potential-risk groups, which further accelerates conducting accurate investigations, prevention, and monitoring [41]. During the pandemic, apps for tracking, tracing, and early warning have been successfully implemented in many countries to control the spread of the coronavirus. Tracking the flow of personnel helped predict the spread trend of the pandemic and improved the efficiency of prevention and control work.

However, ethical and legal issues [42] are raised with the widespread use of contact tracing and monitoring technologies [43]. The pressure to protect personal information is higher than ever. User data is a "disaster area," as privacy can be leaked. Some apps claim that data will be encrypted on the user's mobile phone for a few days and then permanently deleted if users have not been exposed to patients who are infected. Some apps need to obtain the user's consent on two aspects: user's agreement with data being collected while using the app and agreement on releasing relevant personal data once they are found to have contact with patients who are infected.

Data opening and sharing in the context of pandemic prevention and control cannot exceed the reasonable limits. The disclosure and use of data need to protect the rights and interests of citizens to achieve a dynamic balance between public governance and citizen protection. On the one hand, making good use of data technology has been proven to optimize the governance of the public health emergency; on the other hand, personal information and privacy data protection boundaries should be strictly guarded to avoid unnecessary and irreparable damage. Technical tools such as data desensitization and blockchain technologies should be applied along with detailed personal information classification standards, strict privacy regulations, and policies.

Big Data

With the threat of the COVID-19 pandemic, an urgent public health crisis that produces massive data from multiple sources, the use of big data technology can provide the public and decision makers with more complete, continuous, accurate, and timely pandemic prevention information and traceable disease source-based methods [44].

Through big data technology along with geographic location and time stamp information, it is possible to analyze the movement trajectory of affected persons [45]; comprehensively track the movement trajectories of patients who are infected, suspected patients, and related contacts; and accurately describe the cross-regional infiltration. The health care database could be integrated with immigration and customs data to generate real-time alarms during clinic visits to assist in identifying infected cases [46]. Movement tracking has provided powerful data support for the prevention and control of the pandemic.

Conclusion

Currently, the prevention and control of COVID-19 is in a critical period. The construction of integrated intelligent health care systems through novel health technology applications plays a vital role in blocking the spread of the pandemic. The intelligent health care system integrates strategic emerging health technologies and health care delivery services and practices, such as AI, big data, 5G, IoT, cloud computing technology, sensor technology, telehealth service, mobile health apps, and HIE practices. This system will form a new mode of innovation and upgrading of traditional medical and health informatization.

Meanwhile, it is also critical to establish security and privacy protecting systems to enhance the infrastructure, medical data,

and management system. It is important to prevent patient information leakage, create a safer and more convenient medical treatment environment, and provide a strong security guarantee for the intelligent health care system.

Through combining health technology and health care systems, diagnosis efficiency and patients' medical experiences can be improved, and remote sharing of high-quality medical resources and real-time information interaction can also be achieved. The integrated system can effectively alleviate the problems of medical resource shortages, uneven distribution of health care quality, and shortages of health care workers. The establishment of an integrated intelligent health care system for COVID-19 pandemic prevention and control will also provide a positive reference for the design and development of subsequent intelligent health care platforms for other public health crises.

Conflicts of Interest

None declared.

References

1. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020 Mar 28;395(10229):1054-1062 [FREE Full text] [doi: 10.1016/S0140-6736(20)30566-3] [Medline: 32171076]
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, China Novel Coronavirus InvestigatingResearch Team. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020 Feb 20;382(8):727-733 [FREE Full text] [doi: 10.1056/NEJMoa2001017] [Medline: 31978945]
3. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). Int J Surg 2020 Apr;76:71-76 [FREE Full text] [doi: 10.1016/j.ijsu.2020.02.034] [Medline: 32112977]
4. Peeri NC, Shrestha N, Rahman MS, Zaki R, Tan Z, Bibi S, et al. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? Int J Epidemiol 2020 Feb 22 [FREE Full text] [doi: 10.1093/ije/dyaa033] [Medline: 32086938]
5. Kuperman GJ. Health-information exchange: why are we doing it, and what are we doing? J Am Med Inform Assoc 2011;18(5):678-682 [FREE Full text] [doi: 10.1136/amiajnl-2010-000021] [Medline: 21676940]
6. Ammenwerth E, Iller C, Mahler C. IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. BMC Med Inform Decis Mak 2006 Jan 09;6:3 [FREE Full text] [doi: 10.1186/1472-6947-6-3] [Medline: 16401336]
7. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr 2020 Apr 14;14(4):337-339 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.012] [Medline: 32305024]
8. Allam Z, Dey G, Jones DS. Artificial intelligence (AI) provided early detection of the coronavirus (COVID-19) in China and will influence future urban health policy internationally. AI 2020 Apr 13;1(2):156-165. [doi: 10.3390/ai1020009]
9. Naudé W. Artificial Intelligence against COVID-19: an early review. IZA 2020 Apr [FREE Full text]
10. Hao K. This is How the CDC is Trying to Forecast Coronaviruss Spread. MIT Technology Review. 2020 Mar 13. URL: https://www.technologyreview.com/s/615184/the-coronavirus-is-the-first-true-social-media-infodemic/
11. Yang Z, Zeng Z, Wang K, Wong S, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020 Mar;12(3):165-174. [doi: 10.21037/jtd.2020.02.64] [Medline: 32274081]
12. Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals 2020 May 08:109864 [FREE Full text] [doi: 10.1016/j.chaos.2020.109864] [Medline: 32390691]
13. Srinivasa Rao ASR, Vazquez JA. Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine. Infect Control Hosp Epidemiol 2020 Mar 03:1-5 [FREE Full text] [doi: 10.1017/ice.2020.61] [Medline: 32122430]
14. Wang S, Kang B, Ma J, et al. A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). MedRxiv 2020 Apr 24. [doi: 10.1101/2020.02.14.20023028]
15. Mokhtarian P. Defining Telecommuting. UC Davis Institute of Transportation Studies 1991 [FREE Full text]
16. Handy S, Mokhtarian P. The future of telecommuting. Futures 1996 Apr;28(3):227-240 [FREE Full text] [doi: 10.1016/0016-3287(96)00003-1]
17. Perednia DA, Allen A. Telemedicine technology and clinical applications. JAMA 1995 Feb 08;273(6):483-488. [Medline: 7837367]
18. Wootton R. Recent advances: telemedicine. BMJ 2001 Sep 08;323(7312):557-560 [FREE Full text] [doi: 10.1136/bmj.323.7312.557] [Medline: 11546704]
19. Tuckson RV, Edmunds M, Hodgkins ML. Telehealth. N Engl J Med 2017 Oct 19;377(16):1585-1592. [doi: 10.1056/NEJMsr1503323] [Medline: 29045204]

20. Dorsey ER, Topol EJ. State of telehealth. N Engl J Med 2016 Jul 14;375(2):154-161. [doi: 10.1056/NEJMra1601705] [Medline: 27410924]
21. Weinstein RS, Lopez AM, Joseph BA, Erps KA, Holcomb M, Barker GP, et al. Telemedicine, telehealth, and mobile health applications that work: opportunities and barriers. Am J Med 2014 Mar;127(3):183-187. [doi: 10.1016/j.amjmed.2013.09.032] [Medline: 24384059]
22. Eysenbach G. SARS and population health technology. J Med Internet Res 2003;5(2):e14 [FREE Full text] [doi: 10.2196/jmir.5.2.e14] [Medline: 12857670]
23. Hong Z, Li N, Li D, Li J, Li B, Xiong W, et al. Telemedicine during the COVID-19 pandemic: experiences from Western China. J Med Internet Res 2020 May 08;22(5):e19577 [FREE Full text] [doi: 10.2196/19577] [Medline: 32349962]
24. Hollander JE, Carr BG. Virtually perfect? Telemedicine for covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]
25. Tanne JH, Hayasaki E, Zastrow M, Pulla P, Smith P, Rada AG. Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. BMJ 2020 Mar 18;368:m1090. [doi: 10.1136/bmj.m1090] [Medline: 32188598]
26. Elliott T, Yopes MC. Direct-to-consumer telemedicine. J Allergy Clin Immunol Pract 2019;7(8):2546-2552. [doi: 10.1016/j.jaip.2019.06.027] [Medline: 31706486]
27. Portnoy JM, Waller M, De Lurgio S, Dinakar C. Telemedicine is as effective as in-person visits for patients with asthma. Ann Allergy Asthma Immunol 2016 Sep;117(3):241-245. [doi: 10.1016/j.anai.2016.07.012] [Medline: 27613456]
28. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313 [FREE Full text] [doi: 10.1177/1357633X20916567] [Medline: 32196391]
29. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Health Surveill 2020 Apr 02;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]
30. Andrews JG, Buzzi S, Choi W, Hanly SV, Lozano A, Soong ACK, et al. What will 5G be? IEEE J Select Areas Commun 2014 Jun;32(6):1065-1082. [doi: 10.1109/jsac.2014.2328098]
31. Di Ciaula A. Towards 5G communication systems: are there health implications? Int J Hyg Environ Health 2018 Apr;221(3):367-375. [doi: 10.1016/j.ijheh.2018.01.011] [Medline: 29402696]
32. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc 2010;17(3):288-294 [FREE Full text] [doi: 10.1136/jamia.2010.003673] [Medline: 20442146]
33. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114. [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
34. Ashton K. That "Internet of Things" thing. RFID J 2009;22(7):97-114 [FREE Full text]
35. Atzori L, Iera A, Morabito G. The Internet of Things: a survey. Computer Networks 2010 Oct;54(15):2787-2805. [doi: 10.1016/j.comnet.2010.05.010]
36. Perkel JM. The Internet of Things comes to the lab. Nature 2017 Jan 30;542(7639):125-126. [doi: 10.1038/542125a] [Medline: 28150787]
37. Ye J, Li N, Lu Y, Cheng J, Xu Y. A portable urine analyzer based on colorimetric detection. Analytical Methods 2017;9(16):2464-2471. [doi: 10.1039/c7ay00780a]
38. Zhang J, Fu R, Xie L, Li Q, Zhou W, Wang R, et al. A smart device for label-free and real-time detection of gene point mutations based on the high dark phase contrast of vapor condensation. Lab Chip 2015 Oct 07;15(19):3891-3896. [doi: 10.1039/c5lc00488h] [Medline: 26266399]
39. Reeves JJ, Hollandsworth HM, Torriani FJ, Taplitz R, Abeles S, Tai-Seale M, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. J Am Med Inform Assoc 2020 Jun 01;27(6):853-859 [FREE Full text] [doi: 10.1093/jamia/ocaa037] [Medline: 32208481]
40. Abeler J, Bäcker M, Buermeyer U, Zillessen H. COVID-19 contact tracing and data protection can go together. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e19359 [FREE Full text] [doi: 10.2196/19359] [Medline: 32294052]
41. Mahmood S, Hasan K, Colder Carras M, Labrique A. Global preparedness against COVID-19: we must leverage the power of digital health. JMIR Public Health Surveill 2020 Apr 16;6(2):e18980 [FREE Full text] [doi: 10.2196/18980] [Medline: 32297868]
42. Ekong I, Chukwu E, Chukwu M. COVID-19 mobile positioning data contact tracing and patient privacy regulations: exploratory search of global response strategies and the use of digital tools in Nigeria. JMIR Mhealth Uhealth 2020 Apr 27;8(4):e19139 [FREE Full text] [doi: 10.2196/19139] [Medline: 32310817]
43. Cho H, Ippolito D, Yu YW. Contact tracing mobile apps for covid-19: Privacy considerations and related trade-offs [preprint]. arXiv 2020 [FREE Full text]
44. Ye J. Identifying significant practice facilitation delays and barriers in primary care quality improvement. J Am Board Fam Med 2020 [FREE Full text]

45. Chen C, Jyan H, Chien S, Jen H, Hsu C, Lee P, et al. Containing COVID-19 among 627,386 persons in contact with the Diamond Princess cruise ship passengers who disembarked in Taiwan: big data analytics. J Med Internet Res 2020 May 05;22(5):e19540 [FREE Full text] [doi: 10.2196/19540] [Medline: 32353827]
46. Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. JAMA 2020 Mar 03. [doi: 10.1001/jama.2020.3151] [Medline: 32125371]

Abbreviations

AI: artificial intelligence
COVID-19: coronavirus disease
EHR: electronic health record
HIE: health information exchange
HIT: health information technology
IoT: Internet of Things
LSTM: long short-term memory
MERS: Middle East Respiratory Syndrome
PGHD: patient-generated health data
PPE: personal protective equipment
SARS: severe acute respiratory syndrome
SEIR: susceptible-exposed-infectious-removed
2019-nCoV: novel coronavirus

Edited by G Eysenbach; submitted 04.05.20; peer-reviewed by E Chukwu, E Da Silva; comments to author 21.05.20; revised version received 22.05.20; accepted 21.06.20; published 14.07.20.

Please cite as:
Ye J
The Role of Health Technology and Informatics in a Global Public Health Emergency: Practices and Implications From the COVID-19 Pandemic
JMIR Med Inform 2020;8(7):e19866
URL: http://medinform.jmir.org/2020/7/e19866/
doi:10.2196/19866
PMID:32568725

Original Paper

Good News and Bad News About Incentives to Violate the Health Insurance Portability and Accountability Act (HIPAA): Scenario-Based Questionnaire Study

Joana Gaia[1], PhD; Xunyi Wang[2], PhD; Chul Woo Yoo[3], PhD; G Lawrence Sanders[1], PhD

[1]State University of New York at Buffalo, Buffalo, NY, United States

[2]Hankamer School of Business, Baylor University, Waco, TX, United States

[3]Florida Atlantic University, Boca Raton, FL, United States

Corresponding Author:
G Lawrence Sanders, PhD
State University of New York at Buffalo
325G Jacobs
Buffalo, NY, New York
United States
Phone: 1 7166452373
Email: mgtsand@buffalo.edu

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2020/9/e24243/

Abstract

Background: The health care industry has more insider breaches than any other industry. Soon-to-be graduates are the trusted insiders of tomorrow, and their knowledge can be used to compromise organizational security systems.

Objective: The objective of this paper was to identify the role that monetary incentives play in violating the Health Insurance Portability and Accountability Act's (HIPAA) regulations and privacy laws by the next generation of employees. The research model was developed using the economics of crime literature and rational choice theory. The primary research question was whether higher perceptions of being apprehended for violating HIPAA regulations were related to higher requirements for monetary incentives.

Methods: Five scenarios were developed to determine if monetary incentives could be used to influence subjects to illegally obtain health care information and to release that information to individuals and media outlets. The subjects were also asked about the probability of getting caught for violating HIPAA laws. Correlation analysis was used to determine whether higher perceptions of being apprehended for violating HIPAA regulations were related to higher requirements for monetary incentives.

Results: Many of the subjects believed there was a high probability of being caught. Nevertheless, many of them could be incentivized to violate HIPAA laws. In the nursing scenario, 45.9% (240/523) of the participants indicated that there is a price, ranging from US $1000 to over US $10 million, that is acceptable for violating HIPAA laws. In the doctors' scenario, 35.4% (185/523) of the participants indicated that there is a price, ranging from US $1000 to over US $10 million, for violating HIPAA laws. In the insurance agent scenario, 45.1% (236/523) of the participants indicated that there is a price, ranging from US $1000 to over US $10 million, for violating HIPAA laws. When a personal context is involved, the percentages substantially increase. In the scenario where an experimental treatment for the subject's mother is needed, which is not covered by insurance, 78.4% (410/523) of the participants would accept US $100,000 from a media outlet for the medical records of a politician. In the scenario where US $50,000 is needed to obtain medical records about a famous reality star to help a friend in need of emergency medical transportation, 64.6% (338/523) of the participants would accept the money.

Conclusions: A key finding of this study is that individuals perceiving a high probability of being caught are less likely to release private information. However, when the personal context involves a friend or family member, such as a mother, they will probably succumb to the incentive, regardless of the probability of being caught. The key to reducing noncompliance will be to implement organizational procedures and constantly monitor and develop educational and training programs to encourage HIPAA compliance.

(JMIR Med Inform 2020;8(7):e15880) doi:10.2196/15880

KEYWORDS

cyber security; data security; Health Insurance Portability and Accountability Act; motivation; economics of crime; rational choice theory

Introduction

Background

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 introduced legislation for protecting the privacy of personal health information. Although the health care industry in the United States is one of the most regulated industries, compliance with the regulations is variable. In 2017, more than 14.6 million people were affected by data breaches [1]. Cybersecurity reports illustrate that health care data breaches will continue to increase [1-4]. Some of these breaches are simply external malicious attacks, but they are often the result of rent-seeking and illegal behaviors of insiders [5-7]. Verizon's 2018 Data Breach Investigations Report paints a bleak picture of the health care industry in which errors and misuse of data are widespread [8,9]. Health care is the only vertical industry that has more insiders behind breaches: 58% when compared with external actors at 42%. This is probably the reason why the majority of the US population does not trust organizations that share health care information [10-12].

The objective of this study was to identify the role that monetary incentives play in the next generation of employees when it comes to violating HIPAA regulations and privacy laws. These individuals are of particular interest because many will also become trusted insiders, with the knowledge and insight to significantly compromise organizational security systems. The research model was developed using the economics of crime and rational choice theory frameworks to identify situations where employees might engage in illegal breach behavior. Scenarios were developed for 5 situations to determine whether monetary incentives could be used to influence subjects to obtain health care information and to release that information. Approximately 35.4% (185/523) to 45.9% (240/523) of the survey participants indicated that there is a price, ranging from US $1000 to over US $10 million, that is acceptable for violating HIPAA laws. In addition, subjects were also asked about their perceived probability of getting caught for violating HIPAA laws. More than 50.1% (262/523) of the participants indicated that the probability of getting caught was more than 74.9% (392/523). Nevertheless, many of them could still be incentivized to violate HIPAA laws. The correlations between the probability of being apprehended and the level of the monetary incentive required for violating HIPAA ranged from 0.14 to 0.43.

Related Work

Foundation Research on the Economics of Crime

Gary Becker's seminal paper on the market for criminal activity posits that potential criminals examine returns on criminal activity as a function of the probability of getting caught or apprehended and the severity of the punishment [13]. He argued that criminals commit crimes when they perceive the expected benefits from crime would exceed the expected cost of crime. Becker received a Nobel Prize for his research on the economics of crime. Becker's [14] economics of crime model has received more than 1000 citations a year, although it was published in 1968.

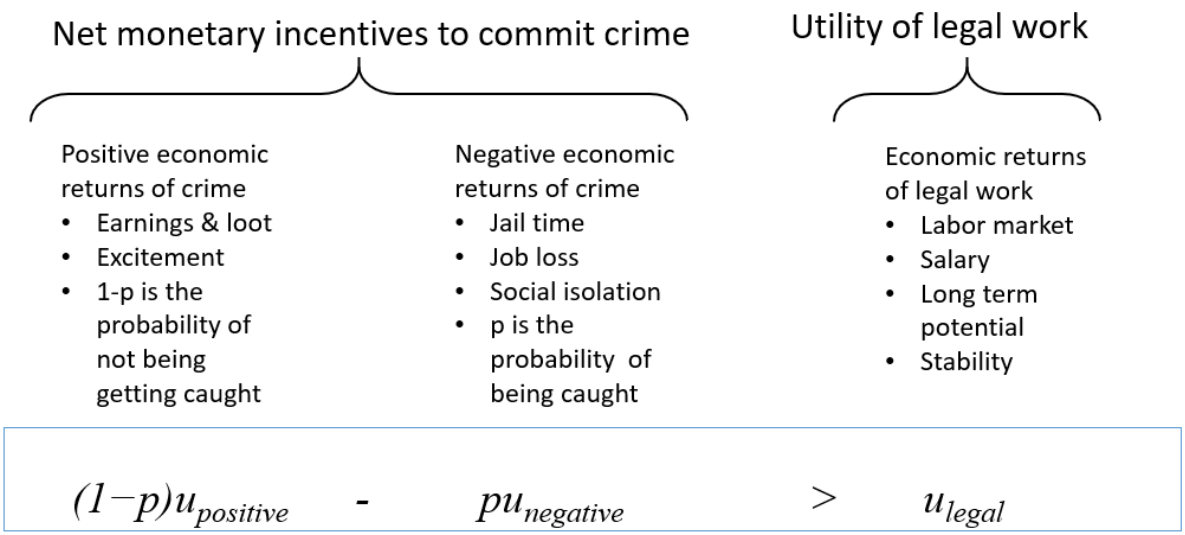
General deterrence theory in the information systems area is used to explore the effects of countermeasures and security policies on protecting information and improving security [15,16]. Early papers by Gopal and Sanders [17,18] examined the role of preventive and deterrent controls on software piracy. Herath and Rao [19] found that the perception of certainty of detections is related to intentions to comply with security policies, but that severity of penalty did not have a deterrent effect. However, deterrence theory research results have been inconsistent and contradictory, and more attention is needed on the theoretical and methodological foundations [20].

General deterrence theory is based on Gary Becker's theory that criminal behavior is deterred when the expected loss (penalty of violating the law) is greater than the expected gain. Many studies involving deterrence theory have focused primarily on the effect of penalties [21]. A framework known as routine activity theory states that a crime can arise from changes in the structured situation or environmental setting, and 4 elements—value, inertia, visibility, and access—would affect the suitability of a target of crime [16,22]. The following paragraphs provide details on the conceptual foundations of the Becker model.

Engaging in criminal activity involves a choice with consequences and opportunities, where individuals perceive them differently. They can be deterred if there is a likelihood of punishment, and the punishment is severe [23]. The market model for crime assumes that offenders, victims, and law enforcement engage in optimizing behavior related to their preferences and that offenders have expectations about returns, the propensity for being caught, and the resulting punishment [23]. This model assumes that potential participants in illegal activities are rational economic actors. Empirical research in the area typically uses an event study that examines whether changes in laws, punishment (incarceration and fines), increases in law enforcement, drug usage, and the economy lead to increases or decreases in criminal activity [24-26].

Wrongdoers use a calculus of rational choice to determine whether to engage in criminal activity [13,27]. An individual will commit a crime if the inequality in Figure 1 holds [28].

Figure 1. The Becker crime utility model.



The $u_{positive}$ term is the expected utility obtained by the potential perpetrator if he or she commits the crime. This utility can mean both monetary and nonmonetary gains. The $u_{negative}$ term is the expected utility resulting from being apprehended and the ensuing punishment. The p term is the probability of being apprehended or getting caught. This is a perception of the risk of offending [27]. The u_{legal} term is the utility derived when he or she does not commit the crime. If the net expected gains from the left side of the inequality are greater than the utility of engaging in legal work on the right side, then the individual will commit the crime.

We illustrate a simplified model of the calculations using 2 equations that form the basis of the model. The criminal will weigh the costs and benefits in the following way:

Benefits=Probability of success × (Gains from crime + Other benefits)

Costs=Probability of getting caught × (Punishment for getting caught + Other costs)

Assume that the expected profits to the potential perpetrator for engaging in illegal activity is US $10,000 and that the probability of success or not getting caught is 90%. The other benefits may be that the potential perpetrator finds excitement from participating and even camaraderie. The utility of these other benefits can be translated into US $2000. Therefore, the total potential benefit is US $10,800 (0.90 × [US $10,000 + US $2000]).

On the costs side, let us assume that the perpetrator perceives that fines of US $16,000 are typically levied as punishment for this type of crime. The other costs might be a loss of job for a few months and social isolation that can be translated into US $6000. The probability of getting caught is 0.10. The total potential cost for engaging in this activity if caught is US $2200 (0.10 × [US $16,000 + US $6,000]).

As the benefits (US $10,800) exceed the costs (US $2200), the individual might engage in criminal activity if this amount of money is perceived as sufficient. As the results of this study show, sometimes there are never enough benefits for people to engage in illegal activities. The other costs are sometimes perceived as being too large, and this translates to a high level of disutility. The other costs could include the loss of a job, prison time, and social desirability effect from a large social network.

There are ongoing discussions and controversy about utility theory and the use of rational decision making among traditional and behavioral economists. Behavioral economists do not abandon the notion that humans can be rational, but they think that there are situations where decision making is less than rational and that more robust models are needed to understand the vagaries of human behavior [29-33]. Our research draws on a combination of traditional economics and behavioral economics to understand the role of incentives in modeling choice behavior related to criminal activity. Empirical evidence supports the role of incentives in terms of labor market experiences and perceptions of the probability of being apprehended and incarcerated [34].

The economics of crime model posits that deterrence will work to counter monetary gains if the penalties are large and if there is a certain level of risk of being caught. There is some empirical evidence that the criminal justice system's ability to deter crime is weaker than thought [26]. However, vibrant labor markets and high manufacturing wages appear to be very effective in deterring crime. In a recent review on the economics of crime, Stephen Levitt of Freakonomics fame [35,36] predicts that there will be fewer research studies on the economics of crime because of declining criminal activity:

> *In some sense, however, public policies to reduce crime (many of them informed by economic thinking) have proven too successful from the perspective of the academic interested in studying crime. With the crime rate at less than half the level it was two decades ago in the United States and lower almost everywhere else in the world as well, the demand for crime research has no doubt also been diminished[37]*

Although it may be true that certain crimes are decreasing, criminal activity involving cybercrime, information security breaches, and privacy intrusions have resulted in substantial dollar losses. HIPAA noncompliance has become a very serious problem. As noted earlier, in 2017, more than 14.6 million people were affected by data breaches, and in the health care industry, errors and misuse of data are widespread [1].

We agree, in part, with Levitt's assertion that academic research has made some gains; however, we believe that the research is at an early stage when it comes to cybercriminal activity, particularly in health care practice. There is evidence that the number of security incidents has decreased, but the dollar amount of financial losses per incident has increased [37]. Underreporting of cybercrime is an elephant-in-the-room problem. Companies are sometimes reticent to report cybercrime because they are embarrassed, and they fear that they will lose customers.

Insider Attacks

Insiders can be current and former employees, contractors, and business partners that have access to an organization's network, system, or data. Insiders can engage in malicious or unintentional activities that negatively affect the confidentiality, integrity, and availability of an organization's information system [38,39].

A recent large-scale, country-wide study found that cyberattacks by outsiders are strategic and often motivated by economic incentives [40]. These attacks can adversely affect business operations and compromise sensitive customer information. However, it appears that trusted insider threats, traced to existing employees, are also related to economic incentives.

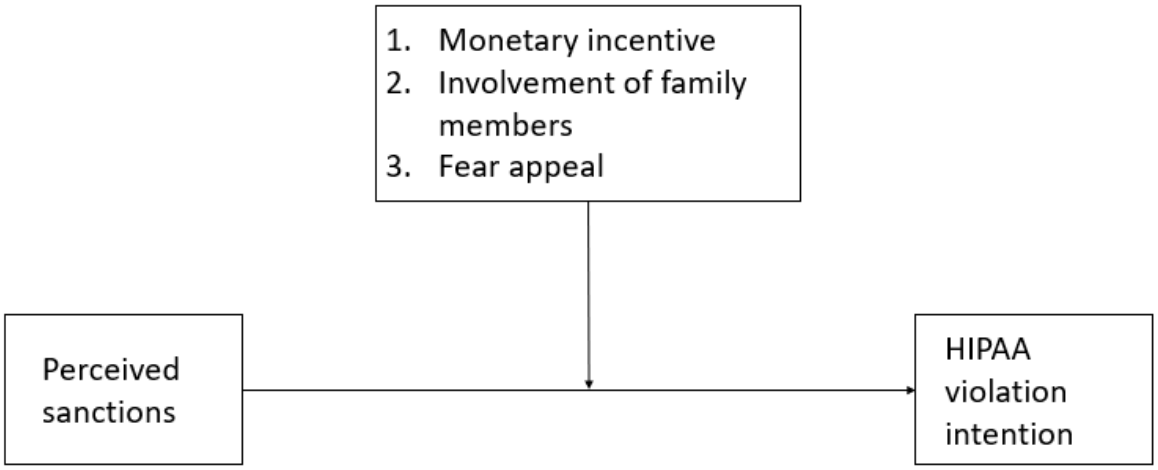
The focus of this research is on insider attacks because they account for a substantial portion of privacy violations, including funds embezzlement; pilfering of trade secrets; theft of customer information and competitive information; and a variety of illegal, fraudulent activities [41], and they can also result in significant losses [42]. Malicious insiders can cause more damage to the organization than traditional hackers [43]. The average cost of an insider attack is US $8 million per year [44], but the fallout from a breach can lead to long-term loss of customers, lawsuits, and damaged reputations.

In some instances, insider security breaches occur because of negligence. For example, some people do not know that they are not supposed to maintain social security numbers in a temporary file or email a medical diagnosis to another doctor without obtaining permission. Insiders pose a considerable threat to organizations as they can bypass several security measures using their knowledge and access to the systems [45]. The motives behind malicious attacks are diverse, including seeking revenge and retribution, thrills, anarchy, and curiosity. Financial motives, however, are the undercurrent of most attacks and include reasons such as student loan debt, financial pressures caused by health care needs or mounting personal debt (eg, credit cards and gambling), or loss of financial stability (job loss or demotion). Threats from trusted insiders are difficult to detect, are embarrassing, damage the reputation of the organization, are often destructive, and cause serious operational disruptions [46].

Hypotheses Development

The primary objective of this study was to identify the role that monetary incentives play in violating HIPAA regulations and privacy laws in the next generation of employees. The conceptual model is presented in Figure 2. The research hypotheses draws on the economics of crime and rational choice theory frameworks to identify situations where employees might engage in illegal breach behavior.

Figure 2. The conceptual model. HIPAA: Health Insurance Portability and Accountability Act.



Our first research hypothesis examines the role of the level of monetary inducements and the perceived probability of being apprehended in violating HIPAA laws.

> *Hypothesis 1: Higher perceptions of being apprehended for violating Health Insurance Portability and Accountability Act regulations are related to higher requirements for monetary incentives.*

Our second research hypothesis focuses on the role of the situational or personal context in violating HIPAA laws. Under the specific context in which a family member or friend needs critical medical assistance that is not covered by insurance, we believe that the relationship will not be as strong as the

relationship in Hypothesis 1. Sometimes, there are compelling personal reasons for committing offenses [41]. They can include medical bills, credit card debt, addictions, and the desire to help a family or friend in need. Scenario 4 involves the need to pay for an experimental operation for the subject's mother. Scenario 5 involves the need to pay for an ambulance airlift for a close friend.

> *Hypothesis 2: Higher Perceptions of Being Apprehended for Violating Health Insurance Portability and Accountability Act Regulations are Related to Higher Requirements for Monetary Incentives When the Personal Context Involves a Family Member or Friend, and the Strength of the Relationship is Not as Strong as in Hypothesis 1.*

The last objective of this study was to determine if the perceived risk or probability of getting caught could be modified by using fear appeals as a deterrent [20]. Approximately 50% of the subjects were targeted to receive information related to real people receiving fines and jail time for violating HIPAA laws (Multimedia Appendix 1). This information is a *fear treatment*, and it is used as a deterrent in this study [47,48].

> *Hypothesis 3: The group receiving the fear appeal treatment will have higher perceptions of the probability of being caught violating HIPAA regulations than the group who did not receive the fear appeals treatment.*

Methods

Participants

The local institutional review board approved the protocol for the pilot study and the main study. A questionnaire was developed to examine the relationships among an individual's propensity to reveal private health care information when offered a monetary incentive and the subject's perception of getting caught violating HIPAA laws. The pilot study involved medical residents and individuals in an executive MBA program, some of who work in the health care industry as executives. After collecting data for the pilot study, significant time was spent in refining the instrument and scenarios to avoid the complexity involved in estimating probabilities and trade-offs found in many research studies involving scenarios and simulated games used to evaluate choice behavior. The data were collected in May 2018.

An important consideration in designing the survey was obtaining information from the subjects on the probability of getting caught if they violated health care regulations. As noted earlier, the questionnaire items were anchored using numerical probabilities and verbal labels because this approach has proven to be a very effective method for eliciting probabilities [49], and it counters some of the measurement problems encountered in measuring perceived arrest rates involved in studies of rational choice theory [50].

The questionnaire was refined and distributed to 574 students in an undergraduate information technology course. This was a voluntary survey, and credits were given for completing the questionnaire. We chose an undergraduate sample because they were more computer proficient, they will be entering the workforce in the immediate future, they are not as aware of HIPAA compliance regulations, and they are less concerned with social desirability issues. These students have majored in business IT, and they have largely been trained for business evaluation and business decision making, but not much on health care, especially the regulations or laws in health care. This is a closed survey that was only open to this particular sample, and we used a password to ensure this.

In social science research, social desirability bias is a type of response bias that is the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. It can take the form of overreporting *good behavior* or underreporting *bad* or undesirable behavior [51]. Social desirability bias occurs when subjects are less prone to answer questions truthfully, which could diminish their social prestige [52]. We assert that the medical interns and the executive MBA participants in the pilot test were deeply concerned with social desirability issues as well as the potential loss of high incomes. That is why we did not revisit that population in the main study. Individuals with high status tend to overreport *good behavior* and underreport *bad behavior*. Social desirability bias is a problem in studies involving abilities, personality, and illegal activities. Subjects with high incomes and status tend to deny illegal acts. In the pilot study, only 6% (6/96) of the participants (3 of the medical residents and 3 of the executive MBAs) succumbed to incentives to violate HIPAA laws. The amount of money required by these individuals ranged from US $50,000 to US $1 billion.

Students in the main study group were given 3 extra points in their final exam for participating in the anonymous survey regardless of completion. We removed subjects with more than 10% (1/10) missing values and subjects who took less than 3 min to complete the survey. The final data set consisted of 523 subjects out of the initial 574 survey participants.

The study subjects consisted of 60% males and 40% females, and their average age was 21 years. The study population consisted of 45% whites, 4% blacks, 4% Hispanics, 45% Asians, and 3% others.

Overview of the Scenarios

Scenarios were adapted from an earlier HIPAA compliance study [53] and redeveloped for 5 situations to determine if monetary incentives could influence subjects to obtain health care information and to release that information to individuals and media outlets (Textbox 1). Multimedia Appendix 2 also illustrates an example of the survey question that elicited a response on how much money a subject would accept to reveal information and their perception of the probability of being apprehended for the nursing scenario. The first 3 scenarios do not incorporate a personal or family situation involved in deciding to accept a monetary incentive. We refer to these scenarios as having no personal context. In addition, we use different income levels in the scenarios, resulting in a differential influence of monetary incentives on insiders' deviant behavior decisions. The salaries in the first 3 scenarios were US $30,000, US $200,000, and US $60,000. We included these salary levels to examine income sensitivities.

Scenarios 4 and 5, however, involve personal or situational contexts. There are situational and transitory influences that lead to committing offenses [41], and they can include financial problems, medical bills, and addictions. Scenario 4 involves the need to pay for an experimental operation for the subject's mother. Scenario 5 involves the need to pay for an ambulance airlift for a close friend. The 2 scenarios incorporate the concept of the value of life and the notion of *identified lives* as distinct from *statistical lives* [30,54]. The economics literature on the value of life distinguishes between identified and statistical lives. Statistical lives have no face or a personal context. Examples include the number of cancer cases that have occurred, the number of lives lost in traffic accidents, and lives lost in natural disasters, whereas identified lives have faces, names, family, and a historical legacy. Identified lives have a personal context.

Textbox 1. Overview of the scenarios.

Scenario 1: Nurse's aide, no personal context

Suppose you are a nurse's aide at a hospital and you earn US $30,000 per year. A friend asks you to get them some information on a patient you have been caring for. What amount of money would you receive to make this acceptable?

Scenario 2: Doctor, no personal context

Suppose you are a doctor at a hospital and you earn US $200,000 per year. A very close friend asks you to access patient information to help them in an upcoming legal battle. What amount of money would you receive to make this acceptable?

Scenario 3: Insurance local celebrity, no personal context

Suppose you work for an insurance company and make US $60,000 per year. A relative asks you to get insurance data on a famous local celebrity from the organization you work for. What amount of money would you receive to make this acceptable?

Scenario 4: Your mother needs an experimental treatment, personal context

Your mother has just been diagnosed with a rare condition that causes kidney failure and is fatal if untreated. This condition can be treated, but the treatment is still considered experimental and is therefore not covered by health insurance, nor is it eligible for any type of financial assistance. The treatment is available both nationally and internationally and costs US $100,000. A media outlet approaches you to get information about a famous politician and offers to pay you US $100,000 for that information. This money can save your mother's life. Would you accept the payment from the media outlet and give the money to your mother?

Scenario 5: Best friend needs air medical transportation, personal context

Your best friend has been in an all-terrain vehicle accident in a rural area of Kansas. He or she has life-threatening injuries and needs air medical transportation to receive lifesaving medical care. The medical air evacuation is not covered by insurance and costs US $50,000. Your best friend will not survive ground transportation or local medical care. A media outlet offers you US $50,000 to obtain the health care records of a famous reality television star. This money can save your best friend's life. Would you accept the payment from the news outlet to give the money to your best friend?

Each scenario also included the following question:

What do you think is the likelihood of getting caught if you accept the money?

Extremely unlikely (0%)

Moderately unlikely (7%)

Slightly unlikely (25%)

Neither likely nor unlikely (50%)

Slightly likely (75%)

Moderately likely (93%)

Extremely likely (100%)

Deterrent Treatment

As noted earlier, we also included a deterrent treatment in the study for half of the participants in the study [19]. The treatment consisted of short vignettes that described instances where individuals received fines and were sentenced for violating HIPAA regulations (Multimedia Appendix 1). Half of the subjects received the treatment.

Research Design

In this section, we provide an overview of the study design. First, 574 students in an undergraduate information technology course voluntarily participated in this survey. Credits were given for completing the questionnaire. Second, half of the participants were given the deterrent treatment, which consisted of short vignettes that described the possible punishments for violating HIPPA regulations. Third, all participants, including both the treated and nontreated ones, completed the survey, where the 5 scenarios were presented. The average completion time was 8.5 min. Therefore, given the clear logic of the survey and the time needed to complete the survey, we believe that survey fatigue is not a serious concern in our study.

Results

Main Findings

We used correlation analysis to explore the relationship between the net monetary incentive to commit a crime and the perceived

probability of being apprehended in Hypothesis 1. Hypothesis 1 was supported. It shows that higher perceptions of being apprehended for violating HIPAA regulations are related to higher requirements for monetary incentives. The correlations between the probability of getting caught and the amount of money that the subjects would accept to provide the information were 0.44 (P<.001) for the nursing scenario, 0.25 (P<.001) for the doctor scenario, and 0.43 (P<.001) for the insurance scenario. Differences in income can explain the differences in the correlations for the nurse/insurance scenarios as compared with the doctor scenario. The nurse aide's salary was US $30,000; the doctor's salary was US $200,000; and the insurance agent' s salary was US $60,000. Referring back to the Becker crime utility model in Figure 1, the monetary incentives to commit a crime on the left side would have to be substantially greater than the utility of legal work on the right side. We had posited that the students would not be aware of HIPPA laws; however, approximately 51% agreed or strongly agreed that they were aware of HIPPA regulations. This variable, however, did not have a statistically significant effect on the results when included in the analysis.

These results provide strong support for Hypothesis 1, showing that higher perceptions of being caught for violating HIPAA regulations are related to higher requirements for monetary incentives. Individuals in the study that perceive higher levels of risk of being caught, in essence, will require more money to participate in an illegal act.

To improve the readability of the instrument crosstabs, we collapsed the amount of money from 11 to 5 categories and the probability of getting caught from 7 to 3 categories. Many of the subjects felt that the probability of getting caught for violating a HIPAA law was very high, greater than 93%. In the nursing scenario, 30% (157/523) of the participants thought the probability of getting caught was greater than 93%, and in the doctor scenario, 50% (261/523) of the participants thought the probability of getting caught was greater than 93%. In the insurance scenario, 39% (204/523) of the participants thought the probability of getting caught was greater than 93%. In the mother scenario, it was 37% (194/523), and in the best friend scenario, it was 38% (199/523). Although many of the individuals in the study believed there was a high probability of being caught, a good number of them could be incentivized to violate HIPAA laws. Tables 1-5 show the results. Figure 3 reflects the general trend of the relationship regarding the amount of money it would take to violate a HIPAA regulation based on the probability of getting caught.

Figure 3. Nursing scenario results.

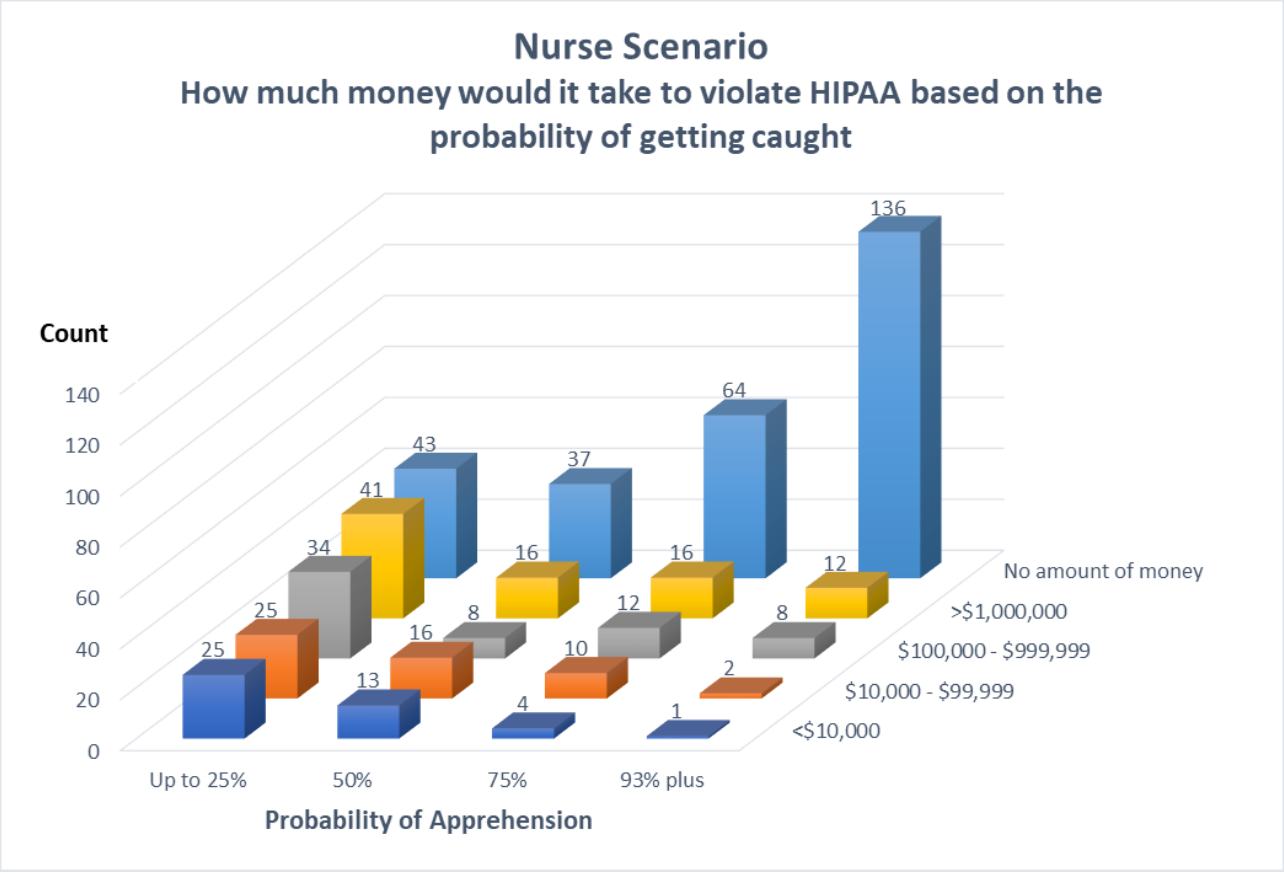


Table 1. Nurse, no personal context (scenario 1).

Scenario 1	Perceived probability of getting caught (R=0.438; *P*<.001; 95% CI 0.36-0.52)				
	≥25%	50%	75%	≤93%	Total, n (%)
Amount of money willing to receive (US $), n					
<10,000	25	13	4	1	43 (8)
10,000-99,999	25	16	10	2	53 (10)
100,000-999,999	34	8	12	8	62 (12)
>1,000,000	41	16	16	12	85 (16)
No amount of money, n	43	37	64	136	280 (54)
Total, n (%)	168 (32)	90 (17)	106 (20)	159 (30)	523 (100)

Table 2. Doctor, no personal context (scenario 2).

Scenario 2	Perceived probability of getting caught (R=0.282; *P*<.001; 95% CI 0.20-0.36)				
	≥25%	50%	75%	≤93%	Total, n (%)
Amount of money willing to receive (US $), n					
<10,000	7	3	5	2	17 (3)
10,000-99,999	9	11	6	9	35 (7)
100,000-999,999	14	9	12	8	43 (8)
>1,000,000	33	12	16	29	90 (17)
No amount of money, n	48	23	52	215	338 (65)
Total, n (%)	111 (21)	58 (11)	91 (17)	263 (50)	523 (100)

Table 3. Insurance company, no personal context (scenario 3).

Scenario 3	Perceived probability of getting caught (R=0.282; *P*<.001; 95% CI 0.20-0.36)				
	≥25%	50%	75%	≤93%	Total, n (%)
Amount of money willing to receive (US $), n					
<10,000	7	3	5	2	17 (3)
10,000-99,999	9	11	6	9	35 (7)
100,000-999,999	14	9	12	8	43 (8)
>1,000,000	33	12	16	29	90 (17)
No amount of money, n	48	23	52	215	338 (65)
Total, n (%)	111 (21)	58 (11)	91 (17)	263 (50)	523 (100)

Table 4. Personal context: your mother needs an experimental treatment (scenario 4).

Scenario 4	Perceived probability of getting caught (R=0.25; *P*<.001; 95% CI 0.17-0.33)				
	≥25%	50%	75%	≤93%	Total, n (%)
Willing to receive US $100,000, n					
No	8	15	22	67	112 (21)
Yes	82	90	114	124	410 (79)
Total, n (%)	90 (17)	105 (20)	136 (26)	191 (37)	522 (100)

Table 5. Personal context: best friend needs air medical transportation (scenario 5).

Scenario 5	Perceived probability of getting caught (R=0.14; *P*<.001; 95% CI 0.05-0.23)				
	≥25%	50%	75%	≤93%	Total, n (%)
Willing to receive US $50,000, n					
No	24	36	34	88	182 (35)
Yes	67	75	87	109	338 (65)
Total, n (%)	91 (18)	111 (21)	121 (23)	197 (38)	520 (100)

The magnitude of the number of individuals who would receive monetary incentives was not expected. We did postulate that there would be some individuals who could be incentivized to violate HIPAA laws, but we thought it would be a small number. In the pilot study, the subjects were medical interns and students enrolled in an executive MBA program. Only 6% (6/96) of the participants (3 medical residents and 3 executive MBAs) succumbed to incentives and violated the HIPAA laws. The amount of money required by these individuals ranged from US $50,000 to US $1 billion. We realize that individuals with high-income potential (medical interns and executive MBAs) would be less prone to violating health care laws, but we did not expect such a dramatic difference.

In the main study, 47.0% (246/523) of the participants received the money in the nursing scenario, 35.0% (183/523) of the participants in the doctor scenario, and 44.9% (235/523) of the participants in the insurance scenario. Again, differences in income might explain the difference, in part. The nurse aide's salary was US $30,000, the doctor's salary was US $200,000, and the insurance agent was US $60,000. Referring back to the Becker crime utility model in Figure 1, the monetary incentives to commit a crime on the left side would have to be substantially greater than the utility of legal work on the right side.

Hypothesis 2 is supported. Recall that it postulates that higher perceptions of being apprehended for violating HIPAA regulations are related to higher requirements for monetary incentives when the personal context involves a family member or friend. However, the strength of this relationship is not as strong as that of the relationship in Hypothesis 1.

Point-biserial correlations are used when there is a dichotomous variable involved. The subjects could answer either a yes or no whether they would accept money to violate a HIPAA regulation. The point-biserial correlation between the probability of getting caught and whether the subjects would accept US $100,000 from a media outlet to pay for an experimental treatment was 0.25 (*P*<.001). The point-biserial correlation between the probability of getting caught and whether the subjects would accept US $50,000 from a media outlet to pay for medical evacuation was 0.14 (*P*=.001).

These correlations are not as strong as those in the first 3 scenarios. The correlations between the probability of getting caught and the amount of money that the subjects would accept to provide the information were 0.44 for the nursing scenario, 0.25 for the doctor scenario, and 0.43 for the insurance scenario.

However, there is more to the story than just the correlations. Looking at the first 3 scenarios, in which there was no personal context, we observed that 47% (246/523) of participants in the nursing scenario indicated that they would be willing to take some level of money to provide patient data, 35% (183/523) of participants in the doctor scenario indicated they would be willing to take some level of money to provide patient data, and 45% (235/523) of participants in the insurance scenario indicated they would be willing to take some level of money to provide insurance data about a celebrity. This is in stark contrast to the 2 personal context scenarios where 79% (413/523) of participants would receive money to save their mothers and 65% (340/523) of participants would receive money to save their best friends.

It is not surprising that 79% of the participants would accept money to save their mother and 65% would accept money to save their best friend. There is a strong *personal motive* to save the lives of individuals who are friends and family, even if there is a strong chance of getting caught. These results are related to how people perceive the difference between *identified lives* and *statistical lives* [30]. Statistical lives involve aggregate numbers, such as 29,000 people die from liver cancer each year. As can be expected, the concepts of statistical life and identified life are very controversial [55]. In the United States, the value of a statistical life has been identified by government agencies to be in the US $7 million range [30]. When a situation involves familiar faces and close relationships with the individual, the use of the statistical value of a life is problematic. It is very difficult to place a value on the life of a family member or close friend. Indeed, the value of a close relative may be infinite. These results support Hypothesis 2 well, which suggests that higher perceptions of being caught violating HIPAA regulations are not related to higher requirements for monetary incentives when the personal context involves a family member or friend.

Prospect theory supports the results for the personal context. Loss of a friend or family member would have a very large impact on an individual's life. The endowment effect also comes into play [56,57]. People value things that they possess, and family and friends are important possessions that are difficult to replace. The endowment construct is related to psychological ownership, and it supports the notion that people overvalue things they perceive they own [58]. Psychological ownership occurs when an individual feels that an object is *theirs* or *mine* [59]. Psychological ownership usually involves some person-object relations. However, it can also be felt toward ideas, words, artistic creations, tablets, phones, people, and virtual avatars [60].

As noted earlier, the situational context matters. In the nursing example, there were 194 individuals in the study that would not

receive any amount of money nor would they turn over patient information to someone. However, those same 194 individuals would take the US $100,000 to pay for an experimental procedure for their mother. The natural question is whether they would take the money because they thought that there would not be a high probability of being caught. However, 124 of the subjects indicated a high probability of getting caught (greater than 93%) but would still help their mother.

Individuals That Are Absolutely Deterred from Violating Health Insurance Portability and Accountability Act Laws

We also counted the number of people who would not violate HIPAA laws at all. There were 14.1% (74/523) of the people in the study that would not receive any money to violate HIPAA regulations for all 5 scenarios. They are what is referred to as absolutely deterred from engaging in criminal behavior. *Absolute deterrence* occurs when individuals refrain from criminal acts because he or she perceives that any level of risk for receiving punishment and the resulting punishment is not acceptable [41,61]. In essence, the severity, certainty, and swiftness of the punishment are not acceptable to absolutely deterred individuals. It was also interesting to note that 14 people would not help their mother but would help their friend. This result is in contrast to the 85 subjects who would help their mother but would not help their friend.

There Is No Treatment Effect

Hypothesis 3 was not supported. Recall that it postulates that the group receiving the fear treatment will have higher perceptions of being caught violating HIPAA regulations than the group who did not receive the fear treatment.

Information related to real people receiving fines and jail time for violating HIPAA laws was received by 50% of the subjects (Multimedia Appendix 1). This information is a fear treatment and is used as a deterrent [47,48]. As noted earlier, the results of studies involving treatment effects for deterrence have been inconsistent and contradictory [20]. Fear appeals use threats in the form of graphics and narrative warnings to modify behavior. The graphics and text illustrated in Multimedia Appendix 1 had little effect on the probability of getting caught. The means between the group receiving the fear appeal treatment and the group who did not receive the treatment were not statistically significant for any of the scenarios. Earlier research on software piracy and MP3 piracy found a modest, yet statistically significant, effect when the subjects were informed about punishment for software and MP3 piracy [17,62]. Sometimes, fear appeals do not work [47,63]. Possible explanations could be that (1) the degree to which an individual perceives information assets as personally relevant is highly subjective, thus potentially marginalizing the impact of the fear appeal, and (2) the conventional fear appeal rhetorical framework is inadequate in providing threat warnings when it is used in the information security context [63]. We included what would be considered as harsh sanctions as a treatment, and there was still no effect.

There is a notion of readiness to commit crimes. Although a large number of participants in the study were attracted to the monetary gains and the need to protect family members and friends, there is a tipping point. In reaching a state of readiness to violate a law, individuals will need to evaluate whether an offense will be a solution to their needs. In other words:

> *It can therefore be predicted that if the expected utility of illegal actions exceeds that of the legal alternatives, an individual will be more likely to decide to engage in a specific crime at a later date (i.e., they will have reached a state of "readiness")[41].*

Information security research needs a major and fundamental shift toward a reconceptualization of deterrence to account for rational forces and restrictive deterrence [41]. One interesting area for research is how potential opportunities to engage in internal computer abuse are shaped by technical skills and the jobs of the insiders. It is also worth considering whether these same employees with the passage of time have been able to contemplate faults in the systems. People in jobs for a long time understand the deficiencies in all aspects of a system, including security flaws. Job movement is one way to deal with this issue, but in the interest of specialization and productivity, moving people around is rarely embraced as a mechanism to increase security.

Discussion

Principal Findings

This study aimed to examine the role that monetary incentives play in violating HIPAA regulations and privacy laws in the next generation of employees. Scenarios were developed for 5 situations to determine whether monetary incentives could influence subjects to obtain health care information and to release that information. Approximately 35% to 46% of the 523 survey participants indicated that there is a price, ranging from US $1000 to over US $10 million, that is acceptable for violating HIPAA laws. In addition, subjects were also asked about their perceived probability of getting caught for violating HIPAA laws. More than 50% of the participants indicated that the probability of getting caught was more than 75%. Nevertheless, many of them could still be incentivized to violate HIPAA laws. The correlations between the probability of being apprehended and the level of the monetary incentive required for violating HIPAA ranged from 0.14 to 0.43.

In the pilot study consisting of 64 medical residents and 32 executive MBA candidates, just 6% (6/96) of the participants would succumb to monetary incentives and violate HIPAA laws. The amount of money required to incentivize medical residents and executives would also be large, ranging from US $50,000 to US $1 billion.

Between 25% and 30% of the subjects in the main study could be incentivized to violate HIPAA laws if they were offered over US $100,000. This is a substantial amount of money, and it is unlikely that such a sum would be offered to trusted insiders to violate privacy laws. The bad news is that although the number of HIPAA privacy breaches detected is declining, the dollar values of losses are escalating.

In general, individuals who perceive that there is a high probability of being caught are less likely to release private

information. The implication is that technology and improvements in organizational processes could increase the perception of the probability of getting caught. The bad news is that approximately 15% of the subjects in the study would receive money, even if there is a 93% or greater chance of being caught.

Moreover, computer knowledge is not necessary because of the availability of *crime as a service*. Third-party providers can be used in cyberattacks [64]. Anyone can hack and attack and become an amateur hacker using simple automated programming tools and distributed denial-of-service–for-hire attacks and by obtaining billions of compromised passwords from the dark web [65]. Trusted insiders could provide the needed entrée for third-party providers of cyberattacks.

Our last finding is that there is a small chance of being caught, and there is an even smaller chance of being convicted. One security expert estimates that for every individual who gets caught, 10,000 people go free and that for every 1 individual who is successfully prosecuted, 100 get off scot-free or just receive a warning [66].

Between April 2003 and July 2018, there were 186,453 health information privacy complaints submitted to the US Department of Health and Human Services [67]. Of these complaints, 37,670 were investigated, resulting in 26,152 (69%) corrective actions. The Office of Civil Rights has imposed civil penalties of US $78,829,182 for just 55 cases. During that same period, the Department of Justice received 688 cases from the Office of Civil Rights for further criminal investigation. It is very difficult to obtain details about the disposition of criminal HIPAA violations. We conducted a search at the Department of Justice [68] using *HIPAA* as a keyword on their website where the Department of Justice has obtained fines and jail time. As illustrated in Multimedia Appendix 3, there were only 11 cases with fines and jail time.

Most of the subjects in our study thought that there was a high probability of being caught for violating HIPAA laws. For example, in the nursing scenario, 30% (157/523) of the participants indicated that there was a 93% or higher chance of getting caught. Clearly, this is not the case. People, even experts, consistently misestimate statistical probabilities, even when there is new contrary evidence.

There Is Often a Price

Our results suggest that many people have a price. It may be a significant amount of money, or it may be a situation where a family member or friend needs critical medical assistance. Monitoring credit reports is a very invasive and controversial practice, but some companies are turning to credit monitoring as a way to counter breaches prompted by financial gain, although several states have taken steps to ban or limit employer access to credit reports.

The results suggest that the subjects in this sample responded rationally to the mother and the best friend scenarios. They just discounted the negative consequences of getting caught, and they attached a very high value to the lives of their mother and best friend. They also acted rationally in the first 3 scenarios. Some people indicate that there was a low probability of getting caught, but many of those people would still not participate in illegal activities. This result may be related to the Black Swan phenomenon [69]. There may be a low probability of getting caught, but the impact of getting caught could have serious long-term consequences and might be perceived, as such, by some individuals. Fines, possible prison time, loss of a job, and difficulty securing a job in the future can result in high monetary costs and social isolation.

Although there are mechanisms for reporting violations, this is still a complex problem. Organizations need to use educational campaigns as well as monitoring and enforcement strategies that strike the proper balance of protecting health care information and protecting the privacy of individuals against inadvertent violation of HIPAA laws.

Our results illustrate the importance of providing both preventive and deterrent information to increase HIPAA compliance [70]. The key will be to implement organizational procedures and constantly monitor and develop educational and training programs that will provide the appropriate frequency and intensity of deterrent information so that employees will not ignore but will embrace HIPAA compliance.

The Challenge Ahead

The protection of personal information is a significant challenge because this information is ubiquitous, and that information has a monetary value. Businesses use this information to target customer segments. Nonprofits use this information to increase the effectiveness of fundraising campaigns. The dark side of the abundance of personal information is that this information can be compromised and retrieved by insiders and external hackers. Insider threats can come from outside infiltrators who become insiders by phishing and social networking attacks. However, they can also come from insider threats, resulting from homegrown malicious employees who intentionally want to compromise a system for profit and for a variety of reasons, including hacktivism and thrill motives. In many instances, breaches occur because of negligence, for example, some people do not know that they are not supposed to maintain social security numbers in a temporary file or email a medical diagnosis to another doctor without obtaining permission.

Our results suggest that there is a high probability that compromises can occur when employees are presented with monetary incentives, given the right context. These results have serious implications because many security breaches are from insiders [42]. Given that the greatest challenge to organizations is insider threats, the results of this study are provocative.

There are some steps that organizations can take to reduce the chance of security breaches. They can use both preventive and deterrent controls to reduce the probability of minor and major events [71]. Preventive controls impede criminal behavior by forcing the perpetrator to deplete resources [17]. Organizations must have preventive controls in place. These preventive controls include sophisticated monitoring systems technologies and constant attention to authentication protocols to prevent unauthorized access to buildings, software, and databases. Organizations usually focus on preventives because preventives can be implemented, and they are under the control of the

organization. This is in contrast to deterrent strategies that focus on the apprehension and punishment of perpetrators as well as on education, legal campaigns, and fear appeals. Developing security education, training, and awareness is always a challenge. The key is to focus continually on health information security awareness [70]. It is not enough to have employees complete a web-based or even an in-person security training class. Employees need to be immersed in security training, receive feedback, and interact socially with other employees on security issues if the training is to be successful [72]. Some organizations are taking very aggressive steps to counter insider threats from malicious employees, negligent users, and infiltrators. They install software that tracks user logins, monitors file and database usage locally and in the cloud, records web activity, and regularly monitors email activity. These systems, in addition to recording activity, can also be used to send out alerts involving unusual behavior by insiders.

Acknowledgments

This study is based upon the work supported by the National Science Foundation under Grant No 1754085.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Treatment.
[DOCX File , 212 KB - medinform_v8i7e15880_app1.docx]

Multimedia Appendix 2
Example of a web-based questionnaire for the nurse aide scenario.
[DOCX File , 116 KB - medinform_v8i7e15880_app2.docx]

Multimedia Appendix 3
Criminal penalties levied by the Department of Justice.
[DOCX File , 15 KB - medinform_v8i7e15880_app3.docx]

References

1. HIPAA Journal. 2017. Largest Healthcare Data Breaches of 2017 URL: https://www.hipaajournal.com/largest-healthcare-data-breaches-2017/ [accessed 2020-05-29]
2. Uwizeyemungu S, Poba-Nzaou P, Cantinotti M. European hospitals' transition toward fully electronic-based systems: do information technology security and privacy practices follow? JMIR Med Inform 2019 Mar 25;7(1):e11211 [FREE Full text] [doi: 10.2196/11211] [Medline: 30907732]
3. Thilakanathan D, Calvo RA, Chen S, Nepal S, Glozier N. Facilitating secure sharing of personal health data in the cloud. JMIR Med Inform 2016 May 27;4(2):e15 [FREE Full text] [doi: 10.2196/medinform.4756] [Medline: 27234691]
4. Bender JL, Cyr AB, Arbuckle L, Ferris LE. Ethics and privacy implications of using the internet and social media to recruit participants for health research: a privacy-by-design framework for online recruitment. J Med Internet Res 2017 Apr 6;19(4):e104 [FREE Full text] [doi: 10.2196/jmir.7029] [Medline: 28385682]
5. Farahmand F, Spafford EH. Understanding insiders: an analysis of risk-taking behavior. Inf Syst Front 2010 Aug 24;15(1):5-15. [doi: 10.1007/s10796-010-9265-x]
6. Schultz EE. A framework for understanding and predicting insider attacks. Comput Sec 2002 Oct;21(6):526-531. [doi: 10.1016/s0167-4048(02)01009-x]
7. Stavrou V, Kandias M, Karoulas G, Gritzalis D. Business Process Modeling for Insider Threat Monitoring and Handling. Cham: Springer International Publishing; 2014.
8. Verizon Enterprise Solutions. 2018. Insights and Resources | Data Breach Investigations Report URL: https://www.verizonenterprise.com/resources/reports/rp_DBIR_2018_Report_execsummary_en_xg.pdf [accessed 2020-05-29]
9. Pal D, Chen T, Zhong S, Khethavath P. Designing an algorithm to preserve privacy for medical record linkage with error-prone data. JMIR Med Inform 2014 Jan 20;2(1):e2 [FREE Full text] [doi: 10.2196/medinform.3090] [Medline: 25600786]
10. Platt JE, Jacobson PD, Kardia SL. Public trust in health information sharing: a measure of system trust. Health Serv Res 2018 Apr;53(2):824-845 [FREE Full text] [doi: 10.1111/1475-6773.12654] [Medline: 28097657]
11. Müthing J, Brüngel R, Friedrich CM. Server-focused security assessment of mobile health apps for popular mobile platforms. J Med Internet Res 2019 Jan 23;21(1):e9818 [FREE Full text] [doi: 10.2196/jmir.9818] [Medline: 30672738]
12. Prochaska MT, Bird A, Chadaga A, Arora VM. Resident use of text messaging for patient care: ease of use or breach of privacy? JMIR Med Inform 2015 Nov 26;3(4):e37 [FREE Full text] [doi: 10.2196/medinform.4797] [Medline: 26611620]

13. Becker GS. Crime and punishment: an economic approach. J Polit Econ 1968 Mar;76(2):169-217. [doi: 10.1086/259394]
14. Becker G. Crime and punishment: an economic approach. Cham: Springer; 1968:a-68.
15. D'Arcy J, Hovav A, Galletta D. User awareness of security countermeasures and its impact on information systems misuse: a deterrence approach. Inf Syst Res 2009 Mar;20(1):79-98. [doi: 10.1287/isre.1070.0160]
16. Wang J, Gupta M, Rao HR. Insider threats in a financial institution: analysis of attack-proneness of information systems applications. MIS Q 2015 Jan 1;39(1):91-112. [doi: 10.25300/misq/2015/39.1.05]
17. Gopal RD, Sanders GL. International software piracy: analysis of key issues and impacts. Inform Syst Res 1998 Dec;9(4):380-397. [doi: 10.1287/isre.9.4.380]
18. Gopal RD, Sanders GL. Preventive and deterrent controls for software piracy. J Manage Inform Syst 2015 Dec 8;13(4):29-47. [doi: 10.1080/07421222.1997.11518141]
19. Herath T, Rao HR. Protection motivation and deterrence: a framework for security policy compliance in organisations. Eur J Inform Syst 2017 Dec 19;18(2):106-125. [doi: 10.1057/ejis.2009.6]
20. D'Arcy J, Herath T. A review and analysis of deterrence theory in the IS security literature: making sense of the disparate findings. Eur J Inform Syst 2017 Dec 19;20(6):643-658. [doi: 10.1057/ejis.2011.23]
21. Willison R, Warkentin M. Beyond deterrence: an expanded view of employee computer abuse. MIS Q 2013 Jan 1;37(1):1-20. [doi: 10.25300/misq/2013/37.1.01]
22. Cohen LE, Felson M. Social change and crime rate trends: a routine activity approach. Am Sociol Rev 1979 Aug;44(4):588-608. [doi: 10.2307/2094589]
23. Myers SL. Estimating the economic model of crime: employment versus punishment effects. Q J Econ 1983 Feb;98(1):157. [doi: 10.2307/1885572]
24. Levitt SD. Understanding why crime fell in the 1990s: four factors that explain the decline and six that do not. J Econ Perspect 2004 Feb;18(1):163-190. [doi: 10.1257/089533004773563485]
25. Ehrlich I. Capital punishment and deterrence: some further thoughts and additional evidence. J Polit Econ 1977 Aug;85(4):741-788. [doi: 10.1086/260598]
26. Cornwell C, Trumbull WN. Estimating the economic model of crime with panel data. Rev Econ Stat 1994 May;76(2):360. [doi: 10.2307/2109893]
27. Loughran TA, Paternoster R, Chalfin A, Wilson T. Can rational choice be considered a general theory of crime? Evidence from individual-level panel data. Criminology 2016 Jan 8;54(1):86-112. [doi: 10.1111/1745-9125.12097]
28. Draca M, Machin S. Crime and economic incentives. Annu Rev Econ 2015 Aug;7(1):389-408. [doi: 10.1146/annurev-economics-080614-115808]
29. Jolls C, Sunstein CR, Thaler R. A behavioral approach to law and economics. Stanford Law Rev 1998 May;50(5):1471-1550. [doi: 10.2307/1229304]
30. Thaler RH. Misbehaving: The Making of Behavioral Economics. New York: WW Norton & Company; Mar 2017:77-81.
31. Thaler RH. Mental accounting and consumer choice. Mark Sci 2008 Jan;27(1):15-25. [doi: 10.1287/mksc.1070.0330]
32. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. Econometrica 1979 Mar;47(2):263. [doi: 10.2307/1914185]
33. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertainty 1992 Oct;5(4):297-323. [doi: 10.1007/bf00122574]
34. Freeman R. The Economics of Crime. In: Handbook of Labor Economics. Amsterdam: Elsevier; 1999.
35. Snijders T, Bosker R. Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. First Edition. New York, USA: William Morrow; 2005.
36. Levitt S, Dubner S. SuperFreakonomics: Global Cooling, Patriotic Prostitutes, and Why Suicide Bombers Should Buy Life Insurance. First Edition. New York, USA: William Morrow; 2009.
37. PricewaterhouseCoopers. 2018. The Global State of Information Security Survey 2018 URL: https://www.pwc.com/us/en/services/consulting/cybersecurity/library/information-security-survey.html#insight1 [accessed 2019-09-07]
38. Theis M, Trzeciak R, Costa D, Moore A, Miller S. Common Sense Guide to Mitigating Insider Threats, Sixth Edition. Technical Report CMU/SEI-2018-TR-. Pittsburgh, Pennsylvania: Software Engineering Institute, Carnegie Mellon University; 2019. URL: https://resources.sei.cmu.edu/asset_files/TechnicalReport/2019_005_001_540647.pdf [accessed 2020-06-09]
39. Rodrigues JJ, de la Torre I, Fernández G, López-Coronado M. Analysis of the security and privacy requirements of cloud-based electronic health records systems. J Med Internet Res 2013 Aug 21;15(8):e186 [FREE Full text] [doi: 10.2196/jmir.2494] [Medline: 23965254]
40. Hui K, Kim SH, Wang Q. Cybercrime deterrence and international legislation: evidence from distributed denial of service attacks. MIS Q 2017 Feb 1;41(2):497-523. [doi: 10.25300/misq/2017/41.2.08]
41. Willison R, Lowry PB, Paternoster R. A tale of two deterrents: considering the role of absolute and restrictive deterrence to inspire new directions in behavioral and organizational security research. J Assoc Inform Syst 2018;19(12):1187-1216. [doi: 10.17705/1jais.00524]
42. Crossler RE, Johnston AC, Lowry PB, Hu Q, Warkentin M, Baskerville R. Future directions for behavioral information security research. Comput Secur 2013 Feb;32:90-101. [doi: 10.1016/j.cose.2012.09.010]

43. Pramanik S, Sankaranarayanan V, Upadhyaya S. Security policies to mitigate insider threat in the document control domain. In: 20th Annual Computer Security Applications Conference: IEEE. 2004 Presented at: 20th Annual Computer Security Applications Conference; December 6-10, 2004; Tuscon, Arizona p. -. [doi: 10.1109/csac.2004.35]
44. Chickowoski E. Dark Reading. 2018. The 6 Worst Insider Attacks of 2018 – So Far URL: https://www.darkreading.com/the-6-worst-insider-attacks-of-2018---so-far/d/d-id/1332183 [accessed 2020-05-29]
45. Software Engineering Institute - Carnegie Mellon University. 2019. Insider Threat URL: https://insights.sei.cmu.edu/insider-threat/ [accessed 2020-05-29]
46. Roy Sarkar K. Assessing insider threats to information security using technical, behavioural and organisational measures. Inf Secur Tech Rep 2010 Aug;15(3):112-133. [doi: 10.1016/j.istr.2010.11.002]
47. Ruiter RA, Kessels LT, Peters GY, Kok G. Sixty years of fear appeal research: current state of the evidence. Int J Psychol 2014 Apr;49(2):63-70. [doi: 10.1002/ijop.12042] [Medline: 24811876]
48. Peters GY, Ruiter RA, Kok G. Threatening communication: a qualitative study of fear appeal effectiveness beliefs among intervention developers, policymakers, politicians, scientists, and advertising professionals. Int J Psychol 2014 Apr;49(2):71-79 [FREE Full text] [doi: 10.1002/ijop.12000] [Medline: 24811877]
49. McGlone MS, Reed AB. Anchoring in the interpretation of probability expressions. J Pragmatics 1998 Dec;30(6):723-733. [doi: 10.1016/s0378-2166(98)00011-3]
50. Pogarsky G, Roche SP, Pickett JT. Heuristics and biases, rational choice, and sanction perceptions. Criminology 2017 Feb 2;55(1):85-111. [doi: 10.1111/1745-9125.12129]
51. Dodou D, de Winter J. Social desirability is the same in offline, online, and paper surveys: a meta-analysis. Comput Hum Behav 2014 Jul;36:487-495. [doi: 10.1016/j.chb.2014.04.005]
52. Akbulut Y, Dönmez O, Dursun OO. Cyberloafing and social desirability bias among students and employees. Comput Hum Behav 2017 Jul;72:87-95. [doi: 10.1016/j.chb.2017.02.043]
53. Basile, Jennifer L. Disseration. University at Buffalo. 2014. An empirical Investigation on Increasing HIPAA Compliance URL: https://ubir.buffalo.edu/xmlui/handle/10477/51211 [accessed 2020-06-09]
54. Hammitt JK, Treich N. Statistical vs identified lives in benefit-cost analysis. J Risk Uncertainty 2007 Jun 21;35(1):45-66. [doi: 10.1007/s11166-007-9015-8]
55. Russell LB. Do we really value identified lives more highly than statistical lives? Med Decis Making 2013 Dec 30;34(5):556-559. [doi: 10.1177/0272989x13512183]
56. Thaler RH. Asymmetric games and the endowment effect. Behav Brain Sci 2010 Feb 4;7(1):117. [doi: 10.1017/s0140525x00026492]
57. Kahneman D, Knetsch JL, Thaler RH. Anomalies: the endowment effect, loss aversion, and status quo bias. J Econ Perspect 1991 Feb;5(1):193-206. [doi: 10.1257/jep.5.1.193]
58. Brasel SA, Gips J. Tablets, touchscreens, and touchpads: how varying touch interfaces trigger psychological ownership and endowment. J Consum Psychol 2014 Apr;24(2):226-233. [doi: 10.1016/j.jcps.2013.10.003]
59. Pierce JL, Kostova T, Dirks KT. The state of psychological ownership: integrating and extending a century of research. Rev Gen Psychol 2003;7(1):84-107. [doi: 10.1037//1089-2680.7.1.84]
60. Moon J, Hossain MD, Sanders GL, Garrity EJ, Jo S. Player commitment to massively multiplayer online role-playing games (MMORPGs): an integrated model. Int J Electron Comm 2014 Dec 8;17(4):7-38. [doi: 10.2753/jec1086-4415170401]
61. Paternoster R. Absolute and restrictive deterrence in a panel of youth: explaining the onset, persistence/desistance, and frequency of delinquent offending. Soc Probl 1989 Jun;36(3):289-309. [doi: 10.1525/sp.1989.36.3.03a00060]
62. Yoo CW, Sanders GL, Rhee C, Choe Y. The effect of deterrence policy in software piracy: cross-cultural analysis between Korea and Vietnam. Inf Dev 2012 Nov 20;30(4):342-357. [doi: 10.1177/0266666912465974]
63. Johnston AC, Warkentin M, Siponen M. An enhanced fear appeal rhetorical framework: leveraging threats to the human asset through sanctioning rhetoric. MIS Q 2015 Jan 1;39(1):113-134. [doi: 10.25300/misq/2015/39.1.06]
64. Reid A. Financial crime in the twenty-first century: the rise of the virtual collar criminal. In: Ryder N, editor. White Collar Crime and Risk: Financial Crime, Corruption and the Financial Crisis. London: Palgrave Macmillan UK; 2018:231-251.
65. Mathews L. Forbes. 2017. File With 1.4 Billion Hacked And Leaked Passwords Found On The Dark Web URL: https://www.forbes.com/sites/leemathews/2017/12/11/billion-hacked-passwords-dark-web/#5991d6f21f2f [accessed 2018-09-17]
66. Grimes RA. CSO. 2016. Why It's So Hard to Prosecute Cyber Criminals URL: https://www.csoonline.com/article/3147398/data-protection/why-its-so-hard-to-prosecute-cyber-criminals.html [accessed 2019-01-02]
67. The US Department of Health and Human Services (HHS). 2019. URL: https://www.hhs.gov/ [accessed 2020-05-29]
68. US Department of Justice. 2019. URL: https://www.justice.gov/ [accessed 2020-05-29]
69. Taleb NN. The Black Swan: The Impact of the Highly Improbable. UK: Penguin UK; 2008:4.
70. Park EH, Kim J, Wiles LL, Park YS. Factors affecting intention to disclose patients' health information. Comput Secur 2019 Nov;87:101340. [doi: 10.1016/j.cose.2018.05.003]
71. Webb J, Ahmad A, Maynard SB, Shanks G. A situation awareness model for information security risk management. Computers Secur 2014 Jul;44:1-15. [doi: 10.1016/j.cose.2014.04.005]

72. Yoo CW, Sanders GL, Cerveny RP. Exploring the influence of flow and psychological ownership on security education, training and awareness effectiveness and security compliance. Decis Support Syst 2018 Apr;108:107-118. [doi: 10.1016/j.dss.2018.02.009]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act

Edited by G Eysenbach; submitted 16.08.19; peer-reviewed by D Paradice, M Chiarini Tremblay; comments to author 18.09.19; revised version received 13.11.19; accepted 14.05.20; published 20.07.20.

Please cite as:
Gaia J, Wang X, Yoo CW, Sanders GL
Good News and Bad News About Incentives to Violate the Health Insurance Portability and Accountability Act (HIPAA): Scenario-Based Questionnaire Study
JMIR Med Inform 2020;8(7):e15880
URL: https://medinform.jmir.org/2020/7/e15880
doi:10.2196/15880
PMID:32706677

Original Paper

Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing

Debbie Rankin[1], BSc, PhD; Michaela Black[1], BSc, PhD; Raymond Bond[2], BSc, PhD; Jonathan Wallace[2], BA, MSc; Maurice Mulvenna[2], BSc, MPhil, PhD; Gorka Epelde[3,4], PhD

[1]School of Computing, Engineering and Intelligent Systems, Ulster University, Derry~Londonderry, United Kingdom

[2]School of Computing, Ulster University, Jordanstown, United Kingdom

[3]Vicomtech Foundation, Basque Research and Technology Alliance, Donostia-San Sebastián, Spain

[4]Biodonostia Health Research Institute, eHealth Group, Donostia-San Sebastián, Spain

Corresponding Author:
Debbie Rankin, BSc, PhD
School of Computing, Engineering and Intelligent Systems
Ulster University
Derry~Londonderry,
United Kingdom
Phone: 44 28 7167 5841 ext 5841
Email: d.rankin1@ulster.ac.uk

Abstract

Background: The exploitation of synthetic data in health care is at an early stage. Synthetic data could unlock the potential within health care datasets that are too sensitive for release. Several synthetic data generators have been developed to date; however, studies evaluating their efficacy and generalizability are scarce.

Objective: This work sets out to understand the difference in performance of supervised machine learning models trained on synthetic data compared with those trained on real data.

Methods: A total of 19 open health datasets were selected for experimental work. Synthetic data were generated using three synthetic data generators that apply classification and regression trees, parametric, and Bayesian network approaches. Real and synthetic data were used (separately) to train five supervised machine learning models: stochastic gradient descent, decision tree, k-nearest neighbors, random forest, and support vector machine. Models were tested only on real data to determine whether a model developed by training on synthetic data can used to accurately classify new, real examples. The impact of statistical disclosure control on model performance was also assessed.

Results: A total of 92% of models trained on synthetic data have lower accuracy than those trained on real data. Tree-based models trained on synthetic data have deviations in accuracy from models trained on real data of 0.177 (18%) to 0.193 (19%), while other models have lower deviations of 0.058 (6%) to 0.072 (7%). The winning classifier when trained and tested on real data versus models trained on synthetic data and tested on real data is the same in 26% (5/19) of cases for classification and regression tree and parametric synthetic data and in 21% (4/19) of cases for Bayesian network-generated synthetic data. Tree-based models perform best with real data and are the winning classifier in 95% (18/19) of cases. This is not the case for models trained on synthetic data. When tree-based models are not considered, the winning classifier for real and synthetic data is matched in 74% (14/19), 53% (10/19), and 68% (13/19) of cases for classification and regression tree, parametric, and Bayesian network synthetic data, respectively. Statistical disclosure control methods did not have a notable impact on data utility.

Conclusions: The results of this study are promising with small decreases in accuracy observed in models trained with synthetic data compared with models trained with real data, where both are tested on real data. Such deviations are expected and manageable. Tree-based classifiers have some sensitivity to synthetic data, and the underlying cause requires further investigation. This study highlights the potential of synthetic data and the need for further evaluation of their robustness. Synthetic data must ensure individual privacy and data utility are preserved in order to instill confidence in health care departments when using such data to inform policy decision-making.

(JMIR Med Inform 2020;8(7):e18910) doi:10.2196/18910

KEYWORDS

synthetic data; supervised machine learning; data utility; health care; decision support; statistical disclosure control; privacy; open data; stochastic gradient descent; decision tree; k-nearest neighbors; random forest; support vector machine

Introduction

Background

National health care departments hold volumes of data on patients and the population, and this information is not being used to its full potential due to valid privacy concerns. Machine learning has the potential to improve decisions and outcomes in health care, but these improvements have yet to be fully realized. The reasons may be related to issues facing many data scientists and researchers in this area: the limited availability of or access to data or the readiness of health care institutions to share data. Privacy concerns over personal data, and in particular health care data, means that although the data exist, they are deemed too sensitive for public release [1], even for research purposes.

One way to overcome the issue of data availability is to use fully synthetic data as an alternative to real data. The exploitation of synthetic data in health care is at an early stage and gaining attention. Synthetic data are simulated from real data by using the underlying statistical properties of the real data to produce synthetic datasets that exhibit these same statistical properties. Synthetic data can represent the population in the original data while avoiding any divulgence of real personal, potentially confidential, and sensitive data. In the case of health-related data, this would ensure that actual patient records are not disclosed thus avoiding governance and confidentiality issues. There are three types of synthetic data: fully synthetic, partially synthetic, and hybrid synthetic. This work considers fully synthetic data that does not contain original data.

Synthetic data can be used in two ways: to augment an existing dataset thus increasing its size, for times when a dataset is unbalanced due to the limited occurrence of an event or when more examples are required [2,3] and to generate a fully synthetic dataset that is representative of the original dataset, for times when data are not available due to their sensitive nature [4]. The latter is considered in this work as a key requirement for health care data sharing.

Traditionally, data perturbation techniques such as data swapping, data masking, cell suppression, and adding noise have been applied to real data to modify and thus protect the data from disclosure prior to releasing it. However, such methods do not eliminate disclosure risk and can impact the utility of the data, particularly if multivariate relationships are not considered [5]. Synthetic data was first proposed by Rubin [6] and Little [7]. Raghunathan et al [8] implemented and extended upon this, pioneering the multiple imputation approach to synthetic data generation, exemplified in a range of studies [9-14]. Reiter [15] then introduced an alternative method of synthesizing data through a nonparametric tree–based technique that uses classification and regression trees (CART). A more recent technique proposes a Bayesian network approach for synthetic data generation [16]. Synthetic data is considered a secure approach for enabling public release of sensitive data as it goes beyond traditional deidentification methods by generating a fake dataset that does not contain any of the original, identifiable information from which it was generated, while retaining the valid statistical properties of the real data. Therefore, the risk of reverse engineering or disclosure of a real person is considered to be unlikely [17].

While a number of synthetic data generators have been developed, empirical evidence of their efficacy has not been fully explored. This work extends a preliminary study [18] and investigates whether fully synthetic data can preserve the hidden complex patterns supervised machine learning can uncover from real data and therefore whether it can be used as a valid alternative to real data when developing eHealth apps and health care policy making solutions. This will be achieved by experimenting with a range of open health care datasets. Synthetic data will be generated using three well-known synthetic data generation techniques. Supervised machine learning algorithms will be used to validate the performance of the synthetic datasets. Statistical disclosure control (SDC) methods that can further decrease the disclosure risk associated with synthetic data will also be considered.

Overview

To inform the viability of the use of synthetic data as a valid and reliable alternative to real data in the health care domain, we will answer the following research questions:

- What is the differential in performance when using synthetic data versus real data for training and testing supervised machine learning models?
- What is the variance of absolute difference of accuracies between machine learning models training on real and synthetic datasets?
- How often does the winning machine learning technique change when training using real data to training using synthetic data?
- What is the impact of SDC (ie, privacy protection) measures on the utility of synthetic data (ie, similarity to real data)?

To answer these questions, 19 open health care datasets containing both categorical and numerical data were selected for experimentation [19]. Synthetic datasets were generated for each dataset using three popular synthetic data generators that apply CART [15,17], parametric [8,17], and Bayesian network [16] approaches to enable a robust comparison of the three synthetic data generation techniques across a broad range of data.

Initially, we analyzed whether the multivariate relationships that exist in the real data were preserved in the synthetic versions of the data for data generated using each of the three synthetic data generation techniques by computing pairwise mutual information scores for each variable pair combination in each dataset [16]. It is important that such relationships are retained when data are synthesized.

To evaluate the utility of synthetic data for machine learning, we then investigated the performance of supervised machine learning models trained on synthetic data and tested on real data compared with models trained on real data and also tested on the real data. This allowed us to determine if a model developed using synthetic data can classify real data examples as accurately and reliably as a model developed using real data. We considered five supervised machine learning models to compare performance and determine if there were differences in robustness across the models. Standard evaluation metrics were computed for models trained on real and synthetic data, for each machine learning model, and for each dataset [20]. The differences in accuracy for models trained on synthetic data versus models trained on real data were computed to analyze the extent to which synthetic data causes a degradation in model performance, if any.

It is pertinent that the optimal machine learning model built using synthetic data matches the optimal machine learning model that would be selected if real data were used in the model training process. This would provide stakeholders in health care with confidence in the use of synthetic data for model development. Thus, we considered how often the best machine learning classifier built using synthetic data matches the best machine learning model built using real data.

Finally, the impact of a number of SDC methods on model performance was assessed. SDC methods seek to further enhance data privacy; however, this can lead to a loss in usefulness of the data [21], and we considered the extent to which performance degradation occurs as a result of SDC.

This large-scale assessment of the reliability of synthetic data when used for supervised machine learning using 19 health care datasets and 3 synthetic data generation techniques provides an important contribution in relation to the trust and confidence that stakeholders in health care can have in synthetic data. We also propose a pipeline to illustrate how synthetic data can potentially fit within the health care provider context. This work demonstrates the promising performance of synthetic data while highlighting its limitations and future work directions to overcome them.

Synthetic Data: Present and Future Use

The validity and disclosure risk associated with synthetic data has been under investigation by the US Census Bureau since 2003 for the purpose of creating public use data from a combination of sensitive data from the Census Bureau's Survey of Income and Program Participation, the Internal Revenue Service's individual lifetime earnings data, and the Social Security Administration's individual benefit data [22,23]. The goal was to enable the release of synthesized person-level records containing personal and financial characteristics from confidential datasets while preserving privacy. Successful results have led to the release of public use synthetic data files. Researchers can have their work validated against the gold standard (real) data by the Census Bureau, thus enabling them to determine the impact of synthetic data on their exploratory analyses and model development and have confidence in their results while also allowing the Census Bureau to continuously improve their synthesis techniques. The public release of this data has provided significant benefit to the research community and general population, enabling more extensive economic policy research to be performed by groups who could not previously access useful data [24-29]. This work led to the release of further synthetic datasets by the Census Bureau. The Synthetic Longitudinal Business Database comprises data from an annual economic census of establishments in the United States [30]. This dataset provides broad access to rich data that supports the research and policy-making communities in business- and employment-related topics. OnTheMap is a tool using synthetic data to provide information on US citizens such as workforce-related maps, demographic profiles, and reports on analyses of information including the location and characteristics of workers living or working in selected areas, the distance and direction totals between residence and employment locations for workers in selected areas, and disaster event information and the impact of such events on workers and employers [31]. Similarly, synthetic data has also been under investigation in the United Kingdom as a means to provide public access to rich data from UK longitudinal studies [32-34] that contain highly sensitive data linking national census data to administrative data for individuals and their families.

These datasets enable researchers to explore data and develop and test code and models outside the secure environment where real data reside with no restrictions while the data owners provide a mechanism where results, code, and models can be validated on behalf of researchers on the real data within the secure environment and feedback provided. This process increases research productivity while ensuring the development of robust and valid models [35].

While synthetic data have been used to accelerate and democratize business and economic policy research [22-35], the process is not currently in use for health care research, an area that could benefit enormously. With advancements in technology, particularly machine learning and artificial intelligence (AI), the potential to develop diagnostic tools for clinicians and data driven decision-making platforms for health policy-makers is ever increasing [36,37]. Such tools require access to health care data, for example, to train AI algorithms and produce models that can identify health conditions and health-related patterns across the population. Currently, it can take a lengthy period of time for researchers to gain access to health care data, a rich and underused resource, due to privacy concerns [38-42]. For example, in the case of the 40-month Meaningful Integration of Data, Analytics, and Services (MIDAS) Project [36,43] developing a data-driven decision-making tool for health care policy makers, it took more than 20 months to obtain access to the required data due to legal and ethical constraints. In addition, a number of important data variables could not made available, which restricted the utility of the platform under development. With the help of synthetic data, such data, with more or all variables included, could have been made available in a matter of weeks, thus providing more time for development and evaluation of the platform. The platform could then have been installed in health care sites more quickly and connected to real data for validation and comparison of performance for synthetic versus real data, enabling performance tweaks to mitigate bias introduced by synthetic

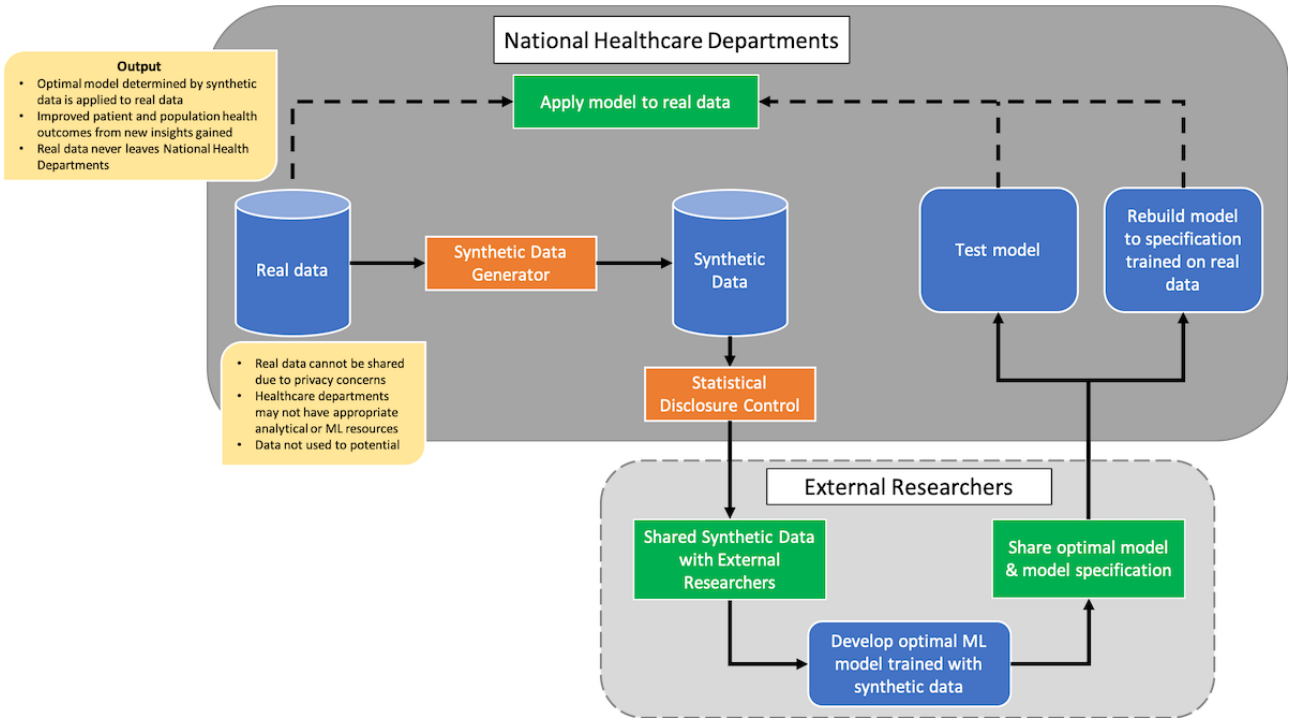
data, if any. Synthetic data could also enable cross-site analytics in various health regions that would enable policy makers to connect their health spaces and potentially provide significant enhancements to cross-national health policy.

The ultimate goal of this work was to further assess the validity and disclosure risk of synthetic data under the stringent conditions associated with health care data with the view to successfully developing a pipeline for use in health care that enables synthetic datasets to be released publicly to researchers, who would otherwise not be able to access the data or access it in a timely fashion, in order to accelerate research by enabling the wider research community to use the data for analysis and model development. The results of such analyses and the models and code developed can then be given to health care departments for validation on the real data and, if effective, put into use by clinicians and health policy-makers.

Synthetic Data Pipeline for Health Care

To understand how health care departments can benefit from synthetic data, we propose the pipeline shown in Figure 1. This is a proposed synthetic data-sharing pipeline provided as an illustration of how synthetic data can potentially work within a real health care setting to expedite data analytics. In future work, we plan to test this pipeline in a real setting. In this pipeline, real data reside within the national health care department infrastructure. The data cannot be shared externally due to the sensitive and private nature. Health care departments may only have a small number of data science staff with the expertise necessary to apply machine learning techniques to many of their datasets, so they cannot maximize the use of their data or discover the potential use of the data due to lack of resources. By applying a synthetic data generation technique to the real data along with SDC measures, a synthetic dataset can be produced and made available to the external research community in place of the real data. External researchers, in large numbers and with wide-ranging expertise, can potentially develop optimal machine learning models trained on the synthetic data and share the performance of the machine learning model, the model itself, and the model specification with the national health care department. The health care department can then test the machine learning model on real data, or in-house technical staff can rebuild the model according to the specification provided by researchers including the program code written by researchers, details of the machine learning algorithm to use (eg, decision tree [DT], support vector machine [SVM]), and the optimal hyperparameter settings determined during development. Using these settings, the model can be rebuilt, this time by training on the real data instead of synthetic data, to which in-house staff have access.

Figure 1. Proposed synthetic data sharing pipeline illustrates how synthetic data could be implemented to expedite health care data analytics.



Methods

Dataset Selection

For experimentation, 19 open health care datasets have been selected from the University of California Irvine Machine Learning Repository [19]. Missing values have been removed from the datasets either by removing features with a high number of missing values or removing observations where a feature contains a missing value. The experimental datasets and their properties are summarized in Table 1. These datasets were selected to enable an analysis of synthetic data performance when applied to datasets of differing volume and data types (categorical and numerical).

Table 1. Summary of experimental datasets.

Dataset and letter designation[a]		Attributes n	Categorical attributes n	Numerical attributes n	Classes/labels n	Observations n
A	Breast Cancer Wisconsin (original)	9	0	9	2	683
B	Breast Cancer	9	9	0	2	277
C	Breast Cancer Coimbra	9	0	9	2	116
D	Breast Tissue	9	0	9	6	106
E	Chronic Kidney Disease	21	12	9	2	209
F	Cardiotocography (3 class)	21	0	21	3	2126
G	Cardiotocography (10 class)	21	0	21	10	2126
H	Dermatology	34	33	1	6	358
I	Diabetic Retinopathy	19	3	16	2	1151
J	Echocardiogram	10	2	8	3	106
K	EEG[b] Eye State	14	0	14	2	14980
L	Heart Disease	13	8	5	2	303
M	Lymphography	18	18	0	4	148
N	Postoperative Patient Data	8	8	0	3	87
O	Primary Tumor	15	15	0	21	336
P	Stroke	10	7	3	2	29072
Q	Thoracic Surgery	16	13	3	2	470
R	Thyroid Disease	22	16	6	28	5786
S	Thyroid Disease (New)	5	0	5	3	215
—	Total	283	144	139	105	58,655

[a]Each dataset has been encoded with a letter (column 1) and will be referenced using this letter for the remainder of the paper.

[b]EEG: electroencephalograph.

Generating Synthetic Data

In this work, we analyzed and assessed the performance of three publicly available synthetic data generation techniques that are based on well known, seminal work in the area [6-10,15,16]: a parametric data synthesis technique, a nonparametric tree-based synthesis technique that uses CART [15], and a synthesis technique that uses Bayesian networks [16]. While other approaches exist, some are developed for specific datasets and problems (eg, SimPop simulates population survey data [44], and Synthea simulates patient population and electronic health record data [45]), whereas these techniques are considered to be more general. The R package Synthpop, developed by Nowak et al [17], provides a publicly available implementation of the parametric- and CART-based synthetic data generators. The DataSynthesizer python implementation, developed by Ping et al [16], provides a publicly available implementation of the Bayesian network-based synthetic data generator. These implementations have been used in this experimental work.

Attributes were synthesized sequentially in both the parametric and CART methods. The synthetic values for the first attribute were synthesized using a random sample from the original observed data since it has no predictors from previously synthesized attributes in the dataset. When synthesizing attributes, both categorical and numerical, with the nonparametric method, the CART method was applied. CART was applied to all variables that had predictors (ie, attributes prior to them in the sequence) and drew from the conditional distributions fitted to the original data using CART models. The parametric method synthesizes attributes based on data type. Numerical attributes were synthesized using normal linear regression. Categorical attributes were synthesized using polytomous logistic regression where the attribute had more than two levels, and logistic regression was applied to synthesize binary categorical variables [17]. The Bayesian network method of synthesizing data learned a differentially private Bayesian network that captured correlation structure between attributes in the real data and drew samples from this model to produce synthetic data [16].

Supervised Machine Learning With Real and Synthetic Data

A key measure of data utility of a synthetic dataset for the purpose of machine learning is to determine how well a supervised machine learning model trained on synthetic data performs when tasked with classifying real data. This determines whether supervised machine learning models will be robust enough to classify real data examples if only synthetic data are provided for the training of these models.

To evaluate whether synthetic datasets could be used as a valid alternative to real datasets in machine learning, for each of the

19 datasets (Table 1), five classification models were trained. Initially, the models were trained and tested on the real data to obtain a performance benchmark. Subsequently, a classifier was trained on each of the synthetic datasets, generated using parametric, CART and Bayesian network techniques, and then tested with the real data. Models were tested on real data only to determine whether a model developed by training on synthetic data can be put into use by health care departments and used to accurately classify new, real examples.

The range of models applied to each dataset were stochastic gradient descent DT, k-nearest neighbors (KNN), random forest (RF), and SVM. This selection of algorithms was applied to determine how well each performed when trained with the real data compared with the synthetic data, with both tested on real data.

The classifiers were implemented using Python's Scikit-Learn 0.21.3 machine learning library and are as follows:

- Stochastic gradient descent classification was implemented using SGDClassifier, a simple linear classifier, with loss="hinge," random_state=0 and all other parameters set to their defaults
- DT classification was implemented using DecisionTreeClassifier, an optimized version of CART, with criterion="gini," max_depth=10, and random_state=0 and all other parameters set to their defaults
- K-nearest neighbors classification was implemented using KNeighborsClassifier with n_neighbors=10, weights="uniform," leaf_size=30, p=2, metric="minkowski," n_jobs=2 and all other parameters set to their defaults
- RF classification was implemented using RandomForestClassifier with criterion="gini," max_depth=10, min_samples_split=2, n_estimators=10, random_state=1 and all other parameters set to their defaults
- SVM classification was implemented using SVC with C=1.0, degree=3, kernel="rbf," probability=True, random_state=None and all other parameters set to their defaults

For training and testing, Python's Scikit-Learn 0.21.3 ShuffleSplit random permutation cross-validator was used with 10 splitting iterations and a train/test split of 75/25. Categorical attributes were transformed into indicator attributes using one-hot encoding.

Statistical Disclosure Control

Synthetic data are considered not to contain real units and therefore the risk of disclosure of a real person is considered to be unlikely [46]. While unlikely, the scenario where some of the generated synthetic data are very similar to the real data resulting in potential disclosure risk must be considered, and where additional protections can be applied to synthetic data, it is recommended to do so. Additional SDC measures beyond data synthesis can be applied as a precautionary measure to add further protections to synthetic data by reducing the risk of reproducing real-person records and replicating outlier data, thus further minimizing the risk of disclosure. There are two broad categories of SDC; rules-based SDC consists of a set of fixed rules governing what data can or cannot be released (eg, a rule setting a specific minimum frequency threshold on a dataset in order for it to be released) and principles-based SDC consists of a broader assessment of risk for a dataset to determine whether it is safe for release (eg, in the case where a specific rule on thresholds may not be applicable because the data cannot be linked back to individuals or in cases where thresholds are not enough to protect individuals from reidentification [47]). SDC measures can be applied, evaluated, and reparameterized as part of the penetration and reidentification testing that health care providers would apply before releasing a synthesized dataset.

The following SDC methods, appropriate for rules-based SDC, have been considered and applied in experimental work to determine their effect on data utility:

- Minimum leaf size (CART method specific): for the CART method, a minimum final leaf node size can be set to avoid the risk of final nodes containing small numbers of records, thus increasing the risk of producing real records (and thus real-person data) in the synthesized data. In SDC experiments, this is set to 10.
- Smoothing: smoothing can be applied to continuous/numerical fields in the synthesized data to reduce the risk of releasing unusual/outlier data. In SDC experiments, gaussian kernel density smoothing is applied to numerical attributes only.
- Unique removal: unique records with variable sequences that are identical to records in the real dataset can be removed. In SDC experiments, this has been applied to synthetic data.

Each of these SDC techniques have been applied to the datasets generated using the CART technique, and the smoothing and unique removal techniques have been applied to datasets generated using the parametric technique. SDC methods have not been applied to data synthesized using the Bayesian network technique.

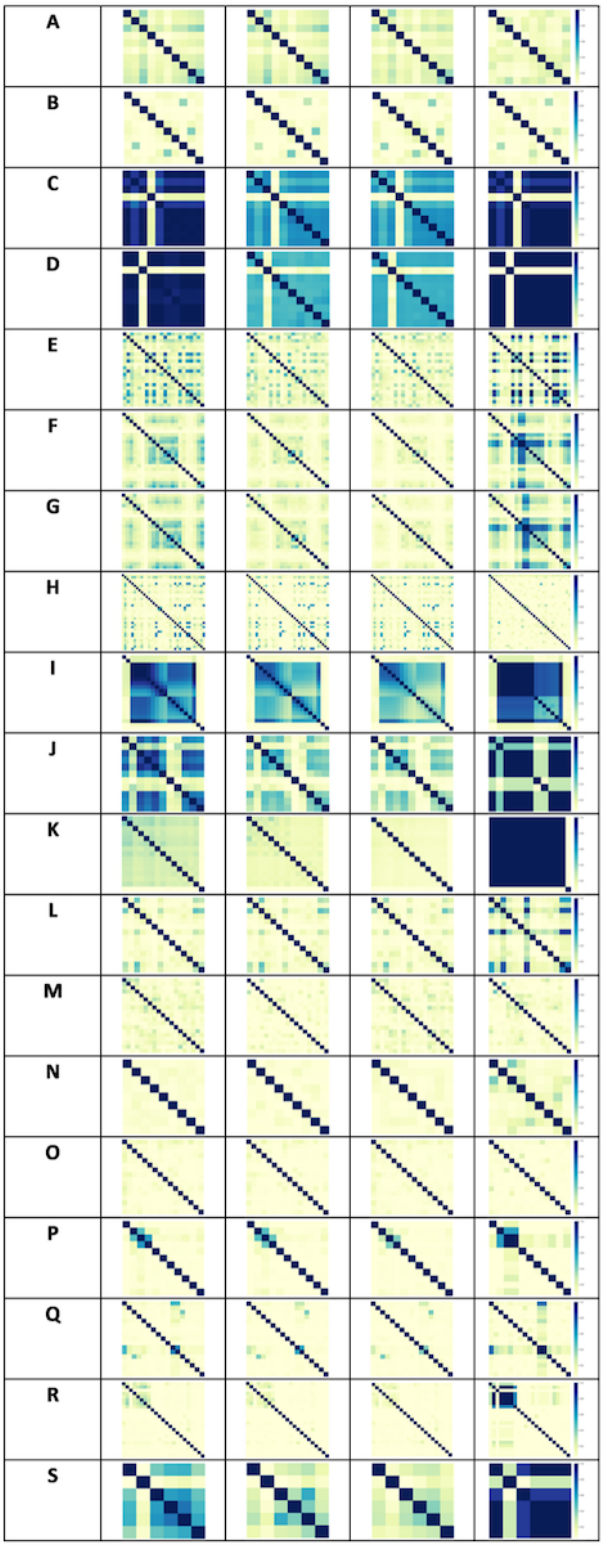
Results

Synthetic Data Properties

Comparison of Variable Relationships

Within a dataset, relationships can exist between variables. When data are synthesized, we wish to determine whether these relationships are preserved and where they are not preserved, whether this relates to the synthesis technique or structure of the dataset. An analysis of these linear relationships was performed by computing the normalized pairwise mutual information score between each pair of attributes. This is a measure of association or similarity where a higher score indicates a greater association between two attributes. Figure 2 provides a visual representation of the normalized pairwise mutual information scores in adjacency heatmaps for each of the 19 datasets (listed in column 1) and enables visual determination of whether the associations found in the real datasets (column 2) are similar to the associations in the synthetic datasets (columns 3-5) for each of the three synthetic data generators.

Figure 2. Pairwise mutual information for the real and synthetic datasets. These adjacency heat maps provide an efficient approach to visually determine whether the associations in the real datasets are similar to the associations in the corresponding synthetic datasets. Column 1 indicates the dataset, columns 2 indicates the pairwise mutual information for the real data, and columns 3-5 indicate the pairwise mutual information for synthetic datasets generated using CART, parametric and Bayesian network approaches, respectively.



The relationships between variables changed slightly in synthetic data generated using the CART and parametric techniques for datasets C-G, I-K, and S, with decreased correlations observed between attribute pairs. These datasets contain mainly and in some cases only numerical attributes. The relationships were largely preserved for the other datasets, which contain mainly and in some cases only categorical attributes, with the exception of dataset A, which contains only numerical attributes.

The relationships between variables also changed slightly in a number of datasets synthesized using the Bayesian network technique (eg, E-G, I-L, N, P-S), with increased correlations observed between attribute pairs. The relationships were largely

preserved in datasets B-D, M, and O, while a slight decrease in correlations between attribute pairs was observed for datasets A and H. In this case, the changes cannot be attributed to a particular data type.

Supervised Machine Learning With Real and Synthetic Data

Performance Comparison

To compare the performance of each model when trained on the synthetic data and tested with the real data, a variety of evaluation metrics were used. The accuracy, precision, recall, and F1 score were computed to determine performance.

The accuracy scores for five machine learning models are shown in Table 2 for datasets A through S. Accuracy scores for models trained on the real data and synthetic data are shown where synthetic data is generated using CART, parametric, and Bayesian network techniques, respectively. The accuracy of the models when trained on synthetic data is lower than the accuracy when trained on real data in 92% (263/285) of cases (ie, machine learning results are less accurate for synthetic data in 92% of cases; Table 3).

Although the accuracy decreases in most cases when using synthetic models, this reduction in accuracy is small. The mean absolute difference in accuracy in models trained with synthetic data across all three synthesizing techniques is lowest for SVM, SGD, and KNN models at 0.058 (6%), 0.064 (6%), and 0.072 (7%), respectively. RF and DT models have larger deviations in accuracy at 0.177 (18%) and 0.193 (19%), respectively (Table 4). This pattern is also consistent when considering results for each of the three synthetic data generators separately. These results are illustrated in the boxplots in Figure 3. The mean absolute difference may provide a reliable indicator of the expected decrease in accuracy in supervised machine learning models when developed using synthetic data. A small yet consistent difference in accuracy is expected and manageable between real and synthetic data.

In addition to accuracy scores, we consider changes to precision, recall, and F1 scores. Precision, recall, and F1 scores decrease in almost all models and for data generated with each synthetic data technique across all 19 datasets (Figure 4). These decreases indicate that the models generated with synthetic data have a higher rate of false-positive and false-negative predictions than models trained with real data. Decreases in precision, recall, and F1 are larger in DT and RF models, consistent with changes in accuracy scores; however, the changes are larger than changes in accuracy for these models. The variance in precision, recall, and F1 differences are also more notable in models trained with synthetic data generated using the Bayesian network approach with less problematic decreases observed in models trained with synthetic data generated using the CART and parametric approaches.

Table 2. Comparison of accuracy scores of five supervised machine learning models trained on real data and synthetic data across 19 datasets. Increase or decrease in accuracy compared with the model trained on real data shown in parentheses.

Dataset and training set[a]	Machine learning algorithm accuracy				
	SGD[b]	DT[c]	KNN[d]	RF[e]	SVM[f]
A					
Real	0.962	1.000 (W[g])	0.975	0.997	0.974
CART[h]	0.966 (+0.004)	0.950 (–0.050)	0.967 (–0.008)	0.965 (–0.032)	0.969 (W) (–0.005)
Parametric	0.932 (–0.030)	0.907 (–0.093)	0.931 (–0.044)	0.927 (–0.070)	0.946 (W) (–0.028)
Bayesian	0.954 (–0.011)	0.924 (–0.076)	0.963 (–0.012)	0.947 (–0.050)	0.967 (W) (–0.007)
B					
Real	0.668	0.931 (W)	0.758	0.924	0.83
CART	0.652 (–0.016)	0.698 (–0.233)	0.765 (+0.007)	0.749 (–0.175)	0.784 (W) (–0.046)
Parametric	0.706 (+0.048)	0.700 (–0.231)	0.748 (–0.010)	0.726 (–0.198)	0.753 (W) (–0.077)
Bayesian	0.674 (+0.006)	0.712 (–0.219)	0.744 (–0.014)	0.741 (–0.183)	0.770 (W) (–0.060)
C					
Real	0.629	1.000 (W)	0.784	0.983	0.905
CART	0.603 (–0.026)	0.652 (–0.348)	0.662 (–0.122)	0.676 (–0.307)	0.729 (W) (–0.176)
Parametric	0.707 (+0.078)	0.702 (–0.298)	0.652 (–0.132)	0.709 (W) (–0.272)	0.700 (–0.205)
Bayesian	0.662 (+0.033)	0.709 (–0.291)	0.664 (–0.144)	0.747 (W) (–0.236)	0.710 (–0.195)
D					
Real	0.632	1.000 (W)	0.726	0.962	0.66
CART	0.502 (–0.130)	0.664 (–0.336)	0.542 (–0.184)	0.706 (W) (–0.254)	0.536 (–0.124)
Parametric	0.472 (–0.160)	0.666 (W) (–0.334)	0.508 (–0.218)	0.628 (–0.334)	0.545 (–0.115)
Bayesian	0.438 (–0.194)	0.592 (–0.408)	0.511 (–0.215)	0.649 (W) (–0.313)	0.557 (–0.103)
E					
Real	0.995	1.000 (W)	0.981	1.000 (W)	0.995
CART	0.972 (–0.023)	0.944 (–0.056)	0.967 (–0.014)	0.995 (W) (–0.005)	0.994 (–0.001)
Parametric	0.964 (–0.031)	0.981 (–0.019)	0.965 (–0.016)	0.988 (W) (–0.012)	0.988 (W) (–0.007)
Bayesian	0.986 (–0.009)	0.957 (–0.043)	0.974 (–0.007)	0.992 (–0.008)	0.993 (W) (–0.002)
F					
Real	0.89	0.985 (W)	0.912	0.982	0.913
CART	0.869 (–0.021)	0.922 (W) (–0.063)	0.883 (–0.029)	0.921 (–0.061)	0.889 (–0.024)
Parametric	0.873 (–0.017)	0.907 (–0.078)	0.886 (–0.026)	0.914 (W) (–0.068)	0.894 (–0.019)
Bayesian	0.880 (–0.010)	0.918 (–0.067)	0.885 (–0.027)	0.924 (W) (–0.058)	0.893 (–0.020)
G					
Real	0.746	0.959	0.78	0.971 (W)	0.82
CART	0.667 (–0.079)	0.848 (W) (–0.111)	0.678 (–0.102)	0.841 (–0.070)	0.748 (–0.072)
Parametric	0.669 (–0.077)	0.805 (W) (–0.154)	0.676 (–0.104)	0.801 (–0.107)	0.737 (–0.083)
Bayesian	0.676 (–0.070)	0.835 (W) (–0.124)	0.676 (–0.104)	0.822 (–0.149)	0.739 (–0.081)
H					
Real	1.000 (W)	0.997	0.98	0.994	0.992
CART	0.940 (–0.060)	0.941 (–0.056)	0.891 (–0.089)	0.958 (W) (–0.036)	0.955 (–0.037)
Parametric	0.935 (–0.065)	0.951 (–0.046)	0.898 (–0.082)	0.959 (W) (–0.135)	0.959 (W) (–0.032)

Dataset and training set[a]	Machine learning algorithm accuracy				
	SGD[b]	DT[c]	KNN[d]	RF[e]	SVM[f]
Bayesian	0.940 (–0.060)	0.952 (–0.045)	0.899 (–0.081)	0.955 (–0.139)	0.959 (W) (–0.032)
I					
Real	0.706	0.845	0.711	0.896 (W)	0.676
CART	0.594 (–0.112)	0.643 (–0.202)	0.634 (–0.077)	0.671 (W) (–0.225)	0.609 (–0.067)
Parametric	0.570 (–0.136)	0.638 (–0.207)	0.624 (–0.087)	0.663 (W) (–0.233)	0.608 (–0.068)
Bayesian	0.609 (–0.097)	0.648 (–0.197)	0.629 (–0.082)	0.667 (W) (–0.229)	0.622 (–0.054)
J					
Real	0.453	0.981 (W)	0.642	0.981 (W)	0.651
CART	0.526 (+0.073)	0.655 (W) (–0.326)	0.579 (–0.063)	0.649 (–0.332)	0.551 (–0.100)
Parametric	0.555 (+0.102)	0.689 (W) (–0.292)	0.606 (–0.036)	0.628 (–0.354)	0.549 (–0.102)
Bayesian	0.545 (+0.092)	0.585 (–0.396)	0.585 (–0.057)	0.602 (W) (–0.379)	0.551 (–0.100)
K					
Real	0.551	0.845	0.864	0.885 (W)	0.551
CART	0.510 (–0.041)	0.531 (W) (–0.314)	0.531 (W) (–0.333)	0.512 (–0.373)	0.531 (W) (–0.020)
Parametric	0.514 (–0.037)	0.545 (W) (–0.300)	0.510 (–0.354)	0.519 (–0.366)	0.531 (–0.020)
Bayesian	0.490 (–0.061)	0.538 (W) (–0.307)	0.510 (–0.354)	0.531 (–0.354)	0.510 (–0.041)
L					
Real	0.851	1.000 (W)	0.861	0.977	0.865
CART	0.791 (–0.060)	0.781 (–0.219)	0.758 (–0.103)	0.803 (W) (–0.174)	0.785 (–0.080)
Parametric	0.822 (W) (–0.029)	0.758 (–0.242)	0.786 (–0.075)	0.809 (–0.168)	0.793 (–0.072)
Bayesian	0.785 (–0.066)	0.738 (–0.262)	0.818 (–0.043)	0.799 (–0.178)	0.834 (W) (–0.031)
M					
Real	0.899	1.000 (W)	0.838	0.986	0.939
CART	0.726 (–0.173)	0.762 (–0.238)	0.762 (–0.076)	0.780 (–0.206)	0.782 (W) (–0.157)
Parametric	0.739 (–0.160)	0.765 (–0.235)	0.757 (–0.081)	0.772 (–0.214)	0.796 (W) (–0.143)
Bayesian	0.681 (–0.218)	0.662 (–0.338)	0.703 (–0.135)	0.746 (–0.240)	0.780 (W) (–0.159)
N					
Real	0.713	0.908 (W)	0.713	0.908 (W)	0.713
CART	0.706 (–0.007)	0.667 (–0.241)	0.715 (+0.002)	0.680 (–0.228)	0.720(W) (+0.007)
Parametric	0.644 (–0.067)	0.614 (–0.294)	0.706 (–0.007)	0.646 (–0.262)	0.708 (W) (–0.005)
Bayesian	0.559 (–0.154)	0.591 (–0.317)	0.706 (W) (–0.007)	0.630 (–0.278)	0.694 (–0.019)
O					
Real	0.449	0.732	0.458	0.762 (W)	0.56
CART	0.338 (–0.111)	0.401 (–0.331)	0.411 (–0.047)	0.410 (–0.352)	0.425 (W) (–0.135)
Parametric	0.317 (–0.192)	0.377 (–0.355)	0.413 (–0.045)	0.397 (–0.365)	0.433 (W) (–0.127)
Bayesian	0.293 (–0.156)	0.336 (–0.396)	0.375 (–0.083)	0.361 (–0.401)	0.419 (W) (–0.141)
P					
Real	0.981	0.985 (W)	0.981	0.982	0.981
CART	0.981 (W) (0.000)	0.977 (–0.008)	0.981 (W) (0.000)	0.981 (W) (–0.001)	0.981 (W) (0.000)
Parametric	0.981 (W) (0.000)	0.976 (–0.009)	0.981 (W) (0.000)	0.981 (W) (–0.001)	0.981 (W) (0.000)
Bayesian	0.981 (W) (0.000)	0.977 (–0.008)	0.981 (W) (0.000)	0.981 (W) (–0.001)	0.981 (W) (0.000)

Dataset and training set[a]	Machine learning algorithm accuracy				
	SGD[b]	DT[c]	KNN[d]	RF[e]	SVM[f]
Q					
Real	0.84	0.932 (W)	0.853	0.928	0.851
CART	0.834 (–0.006)	0.795 (–0.137)	0.850 (–0.003)	0.835 (–0.093)	0.851 (W) (0.000)
Parametric	0.798 (–0.042)	0.811 (–0.121)	0.848 (–0.005)	0.838 (–0.090)	0.849 (W) (–0.002)
Bayesian	0.823 (–0.017)	0.794 (–0.138)	0.846 (–0.007)	0.837 (–0.091)	0.851 (W) (0.000)
R					
Real	0.755	0.989 (W)	0.795	0.961	0.738
CART	0.742 (–0.013)	0.819 (–0.170)	0.761 (–0.034)	0.825 (W) (–0.136)	0.733 (–0.005)
Parametric	0.749 (–0.006)	0.786 (–0.203)	0.764 (–0.031)	0.798 (W) (–0.163)	0.734 (–0.004)
Bayesian	0.748 (–0.007)	0.835 (W) (–0.154)	0.762 (–0.033)	0.832 (–0.129)	0.734 (–0.004)
S					
Real	0.958	1.000 (W)	0.921	1.000 (W)	0.953
CART	0.903 (–0.055)	0.901 (–0.099)	0.899 (–0.022)	0.935 (W) (–0.065)	0.913 (–0.040)
Parametric	0.890 (–0.068)	0.913 (–0.087)	0.912 (–0.009)	0.930 (W) (–0.060)	0.926 (–0.027)
Bayesian	0.905 (–0.053)	0.914 (–0.086)	0.908 (–0.013)	0.936 (W) (–0.064)	0.930 (–0.023)

[a]Training dataset name indicates if real or synthetic data were used to train the model and for synthetic datasets which synthetic data generator was used (ie, CART, parametric, or Bayesian).

[b]SGD: stochastic gradient descent.

[c]DT: decision tree.

[d]KNN: k-nearest neighbors.

[e]RF: random forest.

[f]SVM: support vector machine.

[g](W) highlights the winning classifier for each training set.

[h]CART: classification and regression trees.

Table 3. Changes in accuracy for each machine learning model and synthetic data type (19 datasets and 3 synthetic data generators considered providing 57 synthetic datasets to analyze).

Change in accuracy	Machine learning algorithm					
	SGD[a] (n=57), n (%)	DT[b] (n=57), n (%)	KNN[c] (n=57), n (%)	RF[d] (n=57), n (%)	SVM[e] (n=57), n (%)	Total (n=285), n (%)
Increase	8 (14)	0 (0)	2 (4)	0 (0)	1 (2)	11 (4)
Same	3 (5)	0 (0)	3 (5)	0 (0)	5 (9)	11 (4)
Decrease	46 (81)	57 (100)	52 (91)	57 (100)	51 (89)	263 (92)

[a]SGD: stochastic gradient descent.

[b]DT: decision tree.

[c]KNN: k-nearest neighbors.

[d]RF: random forest.

[e]SVM: support vector machine.

Table 4. Mean absolute difference in accuracy for each machine learning model and synthetic data type.

Synthetic dataset	Mean absolute difference in accuracy per machine learning algorithm				
	SGD[a], n (%)	DT[b], n (%)	KNN[c], n (%)	RF[d], n (%)	SVM[e], n (%)
CART[f]	0.053 (5.3)	0.186 (18.6)	0.069 (6.9)	0.164 (16.4)	0.058 (5.8)
Parametric	0.071 (7.1)	0.189 (18.9)	0.072 (7.2)	0.183 (18.3)	0.060 (6.0)
Bayesian network	0.069 (6.9)	0.204 (20.4)	0.075 (7.5)	0.183 (18.3)	0.056 (5.6)
ALL	0.064 (6.4)	0.193 (19.3)	0.072 (7.2)	0.177 (17.7)	0.058 (5.8)

[a]SGD: stochastic gradient descent.

[b]DT: decision tree.

[c]KNN: k-nearest neighbors.

[d]RF: random forest.

[e]SVM: support vector machine.

[f]CART: classification and regression trees.

Figure 3. Overall change in accuracy for each machine learning model when trained on synthetic data across 19 datasets and 3 synthetic data approaches where classification and regression tree (a), parametric (b), Bayesian network (c), and all approaches combined (d), compared with models trained using real data.

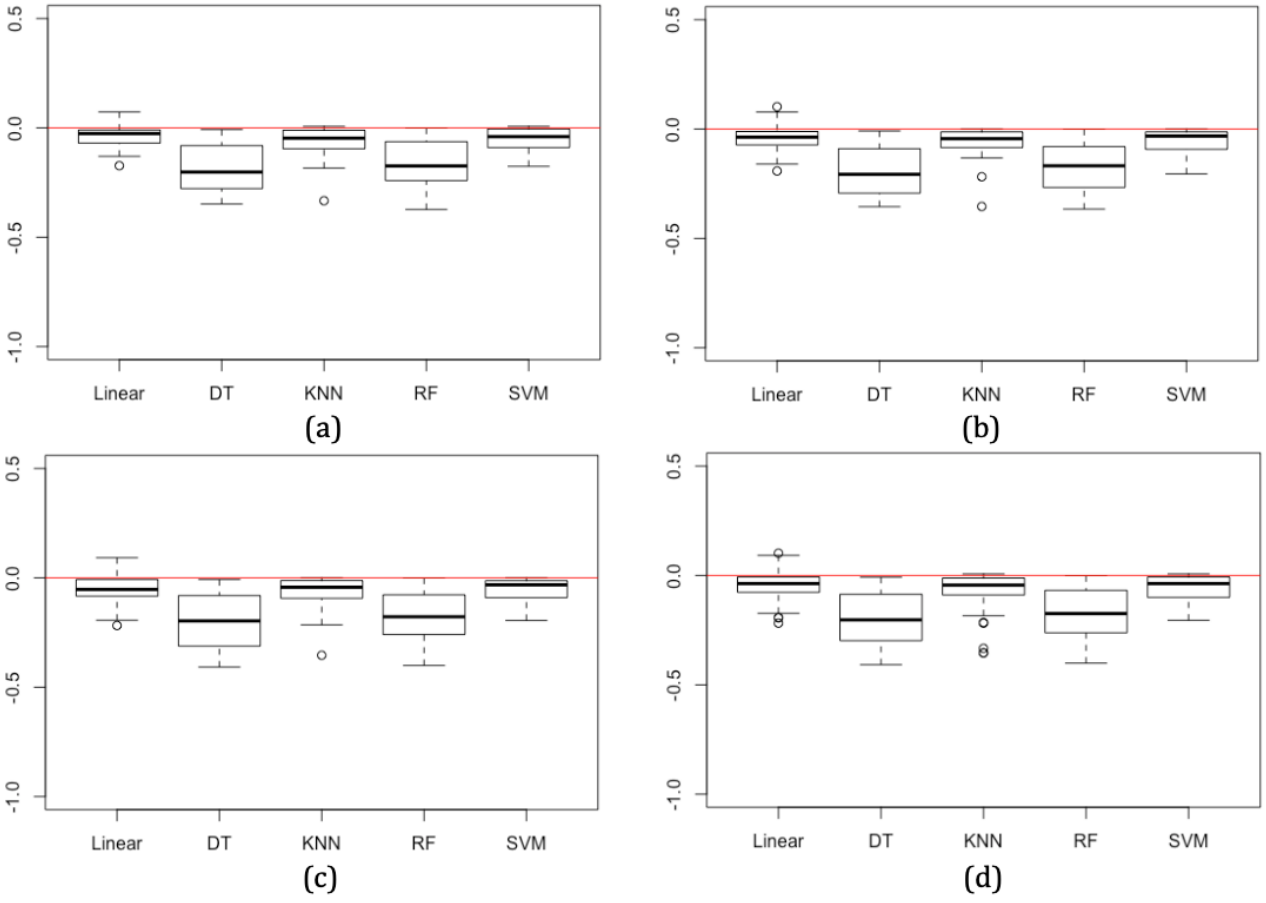
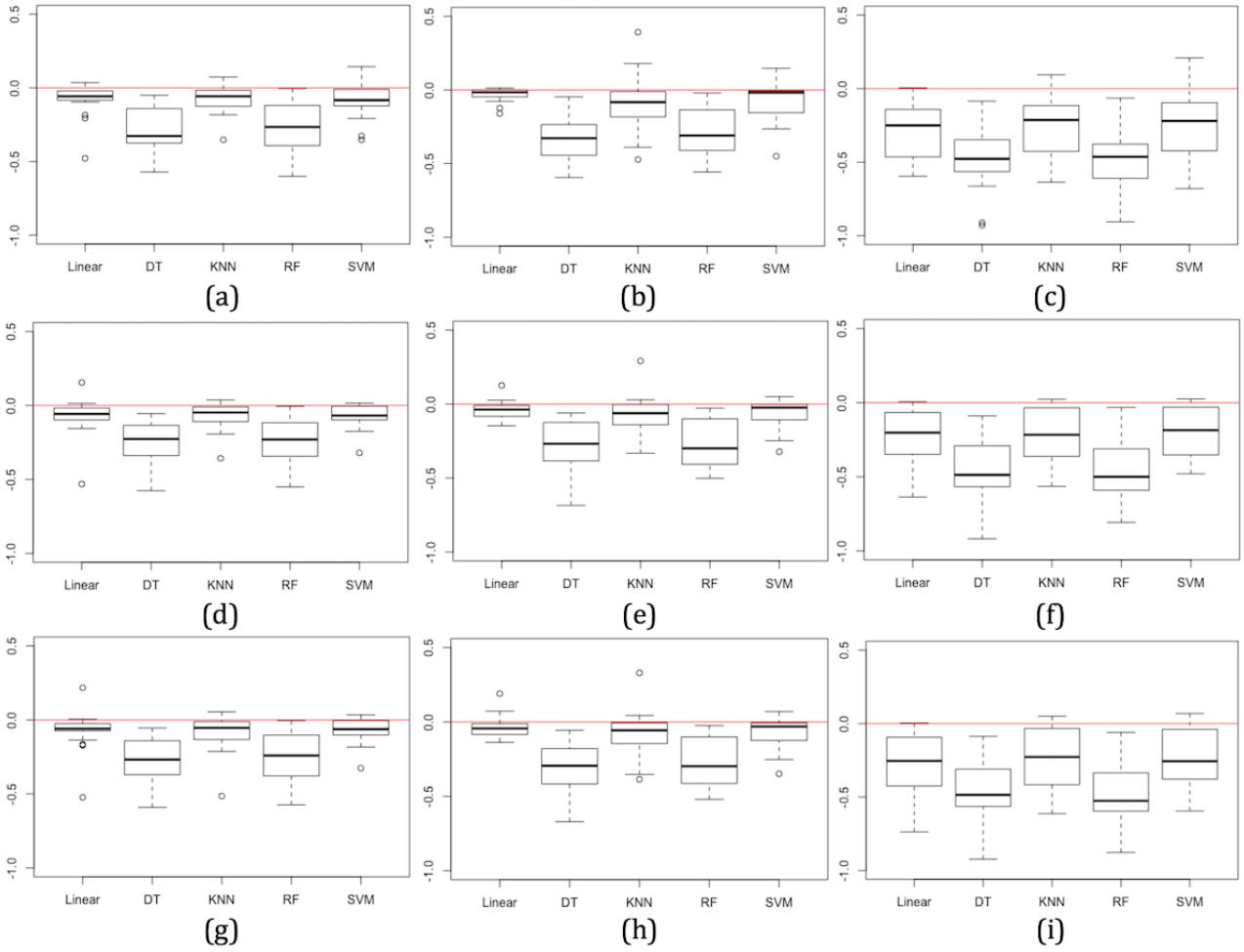


Figure 4. Overall change in precision (a-c), recall (d-f), and F1 (g-i) scores for each machine learning model when trained on synthetic data (from 19 datasets) generated using classification and regression tree (a, d, g), parametric (b, e, h) and Bayesian network (c, f, i) approaches, compared with models trained using real data.



Winning Classifier

In the pipeline described previously, health care departments may wish to release synthetic versions of data to the wider research community for the development of an optimal machine learning model—for example, they may wish to determine the best classifier to use on their real data by making use of the wider range of expertise and scale the external research community can provide. The researchers would be expected to train and test various models and hyperparameters to find the best solution. The researchers would then return a model and/or model specification to the health departments, enabling them to test the model on real data and/or enabling a health department's technical staff to recreate a version of the model, this time trained on the real data to which in-house staff have access. Health departments would have the expectation that this would be the same model determined if real data had been used to develop the best model (ie, it would have been the "winning" model when trained on either synthetic or real data).

We compared the winning classifier when trained and tested on real data with the winning classifier when trained on synthetic data and tested on real data. Table 2 lists the winning classifier (marked as W on each row) for each dataset when trained with real and synthetic data and when tested on the real data.

The winning classifier when trained on real data matches the winning classifier when trained on synthetic data in only 26% (5/19) of cases for synthetic data generated using the CART and parametric methods, and in just 21% (4/19) of cases on data synthesized using the Bayesian network technique (Table 5).

Table 5. Number of instances where the winning classifier trained on synthetic data matches the winning classifier trained on real data across 19 datasets.

Synthetic dataset	Winning classifier matches for real versus synthetic		
	5 classifiers	4 classifiers (DT[a] removed)	3 classifiers (DT and RF[b] removed)
CART[c]	5/19 (26.3)	10/19 (52.6)	14/19 (73.7)
Parametric	5/19 (26.3)	10/19 (52.6)	10/19 (52.6)
Bayesian network	4/19 (21.1)	10/19 (52.6)	13/19 (68.4)
All	14/57	30/57	37/57

[a]DT: decision tree.

[b]RF: random forest.

[c]CART: classification and regression trees.

The DT classifier is most often the winning classifier, in 14/19 datasets, when real data are used to train and test the model, but DT is not the best classifier on synthetic data, winning in only 11/57 cases (Table 2). Tree-based methods (DT and RF) are the winning classifier on real data in 18/19 cases (95%). If we remove DTs from this analysis, the cases where the winning classifier when trained on synthetic data matches the winning classifier when trained on real data almost doubles, increasing to 53% (10/19) of cases for synthetic data generated using each of the three synthesizing techniques (Table 5).

With DTs removed, RF models are now the most frequent winners (18/19) when real data are used to train and test the model. In this case, RF models produce the winning classifier in 32/57 cases (Table 2). If we further remove RFs from this analysis and do not consider tree-based classifiers, cases where the winning classifier when trained on synthetic data matches the winning classifier when trained on real data increases from 53% to 74% (14/19) and 68% (13/19) for data synthesized using CART and Bayesian network techniques, respectively, and remains unchanged for data generated using the parametric technique (Table 5).

A chi-square test is applied with the following null and alternative hypotheses:

- H0: the number of winning classifier matches is equal across all sets of classifiers.
- H1: the number of winning classifier matches increases when DT and RF classifiers are removed.

The level of significance adopted for hypothesis testing is .05 for all tests performed.

The null hypothesis is rejected when the tree-based models (DTs and RFs) are removed (ie, from 5 to 3 classifiers) for data synthesized using the CART and Bayesian network methods (Table 6). Therefore, a significant difference in the matching winning classifiers is observed when tree-based classifiers are removed for these two synthesizing techniques. The null hypothesis could not be rejected in all other cases.

Table 6. Results of chi-square analysis of the difference in matching winning classifiers for models trained on real versus synthetic data.

Synthetic dataset	Winning classifier matches for real versus synthetic		
	5 classifiers	4 classifiers (DT[a] removed)	3 classifiers (DT and RF[b] removed)
CART[c]	0.1843	0.3130	0.0094
Parametric	0.1843	1.0000	0.1843
Bayesian network	0.0927	0.0927	0.0091

[a]DT: decision tree.

[b]RF: random forest.

[c]CART: classification ans regression trees.

Impact of Statistical Disclosure Control

The impact of SDC methods on data utility is considered across all datasets. Table 7 illustrates the effect on model accuracy of applying smoothing (numeric attributes only), removal of unique records, and limiting the minimum leaf size (CART models only) to all synthetic datasets where each method is applicable.

Table 7. Changes in accuracy for each machine learning model and with statistical disclosure control applied.

Change in accuracy	Machine learning algorithm change in accuracy with SDC[a]					
	SGD[b]	DT[c]	KNN[d]	RF[e]	SVM[f]	Total
Smoothing (n=150)						
Increase	4/30 (13.3)	0/30 (0.0)	1/30 (3.3)	0/30 (0.0)	2/30 (6.7)	7/150 (4.7)
Same	2/30 (6.7)	0/30 (0.0)	2/30 (6.7)	0/30 (0.0)	3/30 (10.0)	7/150 (4.7)
Decrease	24/30 (80.0)	30/30 (100.0)	27/30 (90.0)	30/30 (100.0)	25/30 (83.3)	136/150 (90.7)
Unique removal (n=190)						
Increase	4/38 (10.5)	0/38 (0.0)	1/38 (2.6)	0/38 (0.0)	2/38 (5.3)	7/190 (3.7)
Same	2/38 (5.3)	0/38 (0.0)	2/38 (5.3)	0/38 (0.0)	4/38 (10.5)	8/190 (4.2)
Decrease	32/38 (84.2)	38/38 (100.0)	35/38 (92.1)	38/38 (100.0)	32/38 (84.2)	175/190 (92.1)
Minimum leaf size (n=95)						
Increase	2/19 (10.5)	0/19 (0.0)	0/19 (0.0)	0/19 (0.0)	1/19 (5.3)	3/95 (3.2)
Same	1/19 (5.3)	0/19 (0.0)	1/19 (5.3)	0/19 (0.0)	2/19 (10.5)	4/95 (4.2)
Decrease	16/19 (84.2)	19/19 (100.0)	18/19 (94.7)	19/19 (100.0)	16/19 (84.2)	88/95 (92.6)
All (n=435)						
Increase	10/87 (11.5)	0/87 (0.0)	2/87 (2.3)	0/87 (0.0)	5/87 (5.7)	17/435 (3.9)
Same	5/87 (5.7)	0/87 (0.0)	5/87 (5.7)	0/87 (0.0)	9/87 (10.3)	19/435 (4.4)
Decrease	72/87 (82.8)	87/87 (100.0)	80/87 (92.0)	87/87 (100.0)	73/87 (83.9)	399/435 (91.7)

[a]SDC: statistical disclosure control. Each of the 3 types of SDC applied (smoothing, unique removal and minimum leaf size for CART). SDC applied to parametric and CART methods only. Smoothing applied to datasets with numeric attributes only. Minimum leaf size for CART is applicable to CART only.

[b]SGD: stochastic gradient descent.

[c]DT: decision tree.

[d]KNN: k-nearest neighbors.

[e]RF: random forest.

[f]SVM: support vector machine.

In most cases, the machine learning model accuracy decreases when SDC measures are applied to the synthetic data used to train the models. Decreases in accuracy are observed in all DT and RF models and in 83% (72/87), 92% (80/87), and 84% (73/87) of SGD, KNN, and SVM models, respectively. In a small number of cases across SGD, KNN, and SVM models trained on synthetic data with SDC measures applied, no change or a slight increase in accuracy compared with models trained on real data with no SDC measures applied was observed.

The mean absolute difference in accuracy when SDC measures are applied to the training data (compared with machine learning models trained on real data) is small across all machine learning models and for all SDC techniques (Table 8). DT and RF models have the largest difference in accuracy, consistent with earlier results of these models trained on synthetic data with no SDC measures applied. The accuracy decreases are consistent across each SDC measure with no SDC measure affecting data utility more notably than any other. These results are also illustrated in the boxplots in Figure 5. Precision, recall, and F1 scores are also consistent with earlier results when no SDC measures are applied. We therefore consider that the SDC techniques investigated do not have a notable impact on data utility beyond what the standard synthesizers have.

Table 8. Mean absolute difference in accuracy for each machine learning model and statistical disclosure control type.

SDC[a] applied to synthetic dataset	Average change in accuracy per machine learning algorithm				
	SGD[b]	DT[c]	KNN[d]	RF[e]	SVM[f]
Smoothing	0.059 (5.9)	0.190 (19.0)	0.094 (9.4)	0.177 (17.7)	0.060 (6.0)
Unique removal	0.052 (5.2)	0.206 (20.6)	0.072 (7.2)	0.184 (18.4)	0.056 (5.6)
Minimum leaf size	0.061 (6.1)	0.200 (20.0)	0.068 (6.8)	0.180 (18.0)	0.053 (5.3)
All	0.056 (5.6)	0.199 (19.9)	0.078 (7.8)	0.180 (18.0)	0.057 (5.7)

[a]SDC: statistical disclosure control.

[b]SGD: stochastic gradient descent.

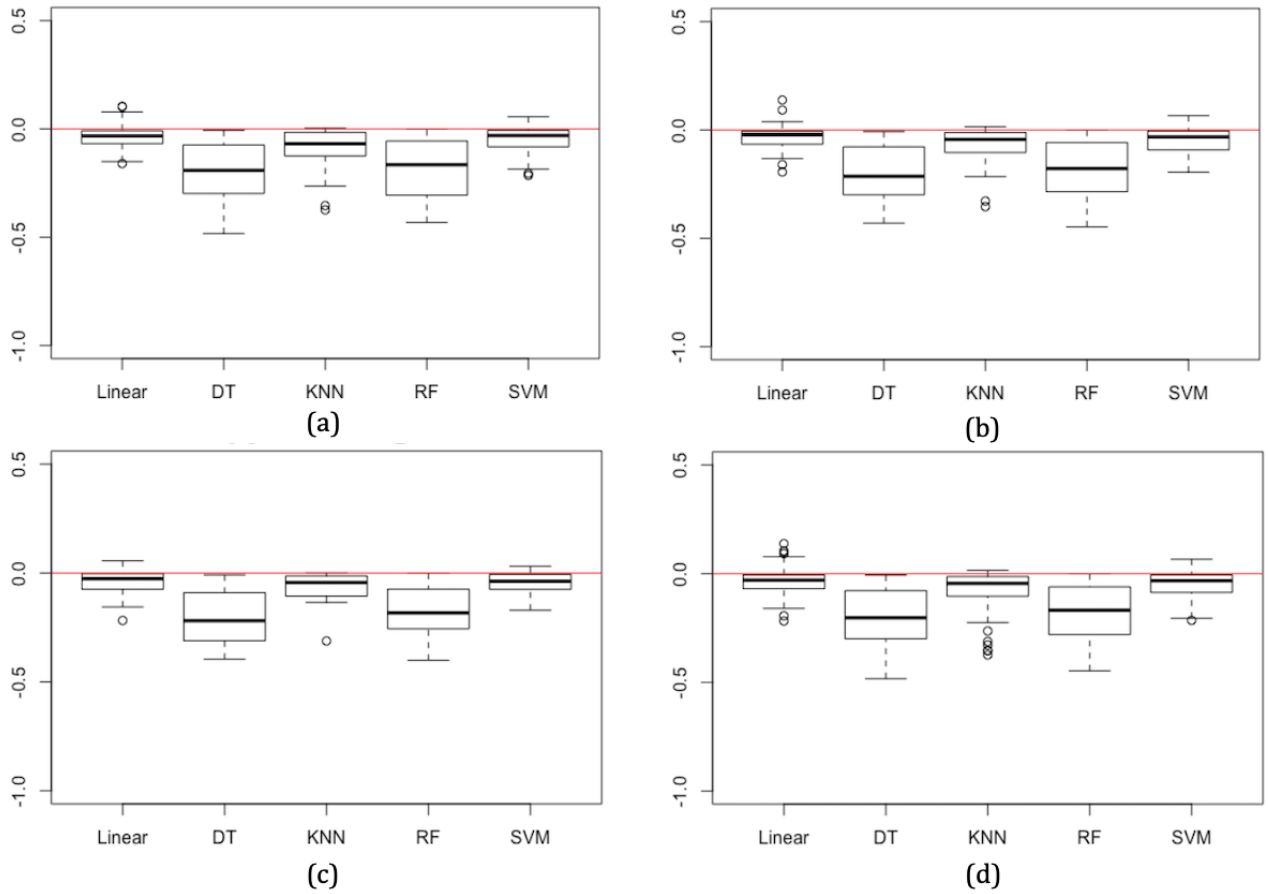
[c]DT: decision tree.

[d]KNN: k-nearest neighbors.

[e]RF: random forest.

[f]SVM: support vector machine.

Figure 5. Overall change in accuracy for each machine learning model when trained on synthetic data across 19 datasets and 2 synthetic data approaches (classification and regression tree [CART] and parametric) and with statistical disclosure control measures applied where smoothing (a; numeric attributes only), unique removal (b), minimum leaf size constrained (c; for CART synthesizer only), and all approaches combined (d), compared with models trained using real data.



We also compare the winning classifier when trained on real data with the winning classifier when trained on synthetic data with SDC applied (Table 9). The winning classifier when trained on synthetic data with SDC applied matches the winning classifier when trained on synthetic data in only 25% (22/87) of cases, consistent with earlier results when SDC measures are not applied. Similar results are observed when each SDC measure is considered individually with the winning classifier matching in models trained with real data compared with models trained using synthetic data with SDC measures of smoothing, unique removal, and minimum leaf size in 27% (8/30), 24% (9/38), and 26% (5/19) of cases, respectively.

Consistent with results in the previous section where SDC measures were not applied, removing tree-based classifiers (DT and RF) from the analysis increases the matches in winning classifiers trained on real compared with synthetic data by 13.3, 36.8, and 36.9 percentage points for each of the SDC measures

of smoothing, unique removal, and minimum leaf size, respectively. Overall, an increase of 28.7 percentage points is observed for all SDC measures when tree-based classifiers are removed.

Table 9. Number of instances where the winning classifier trained on synthetic data with statistical disclosure control applied matches the winning classifier trained on real data across 19 datasets.

Synthetic dataset	Winning classifier matches for real versus synthetic		
	5 classifiers, n (%)	4 classifiers (DT[a] removed), n (%)	3 classifiers (DT and RF[b] removed), n (%)
Smoothing	8/30 (27)	14/30 (47)	12/30 (40)
Unique removal	9/38 (24)	15/38 (40)	23/38 (61)
Minimum leaf size	5/19 (26)	7/19 (37)	12/19 (63)
All	22/87 (25)	36/87 (41)	47/87 (54)

[a]DT: decision tree.

[b]RF: random forest.

A chi-square test is applied with the following null and alternative hypotheses:

- H0: the number of winning classifier matches is equal across all sets of classifiers where SDC measures are applied.
- H1: the number of winning classifier matches increases when DT and RF classifiers are removed where SDC measures are applied.

The level of significance adopted for hypothesis testing is .05 for all tests performed (α=.05).

The null hypothesis is rejected when the tree-based models (DTs and RFs) are removed (ie, from 5 to 3 classifiers) for data synthesized with the SDC measure of unique removal applied (Table 10). Therefore, a significant difference in the matching winning classifiers is observed when tree-based classifiers are removed for this SDC measure. The null hypothesis could not be rejected in all other cases.

Table 10. Results of chi-square analysis of the difference in matching winning classifiers for models trained on real versus synthetic data with statistical disclosure control applied.

Synthetic dataset	*P* values		
	5 classifiers	4 classifiers (DT[a] removed)	3 classifiers (DT and RF[b] removed)
Smoothing	.18	.79	.41
Unique removal	.22	.11	.003
Minimum leaf size	.73	.19	.05

[a]DT: decision tree.

[b]RF: random forest.

Discussion

Principal Findings

The need for synthetic data, particularly in the health care domain, is gaining increasing attention as privacy protection mechanisms are increasingly failing to protect modern data. Due to valid privacy concerns, it is often difficult or impossible to release real health care data thus impeding critical machine learning research that can make use of this data to drive improved patient outcomes and health policy decision-making. Synthetic data has the potential to overcome data availability issues, providing a valid alternative to real data. A small number of synthetic data generators have been proposed in the literature; however, evidence of their efficacy across a large number of datasets and for use in machine learning is thin on the ground.

This work has explored the use of fully synthetic data across 19 health care datasets. Three well-known synthetic data generators have been considered where data is generated using CART, parametric, and Bayesian network techniques. A number of research questions have been answered.

What Is the Differential in Performance When Using Synthetic Data Versus Real Data for Training and Testing Supervised Machine Learning Models?

Compared with models trained and tested on real data, almost all machine learning models have a slightly lower accuracy when trained on synthetic data and tested on real data across all synthesizers and for all machine learning models analyzed; however, the average decrease in accuracy was small in all cases. Although still small, DT and RF models had a larger decrease and variance in accuracy than SGD, KNN, and SVM models. In addition to accuracy, an analysis of precision, recall, and F1 scores also showed decreases in scores in models trained with synthetic data, with Bayesian network-generated data resulting in more variance than data generated using CART and parametric techniques.

What Is the Variance of Absolute Difference of Accuracies Between Machine Learning Models Training on Real and Synthetic Datasets?

The mean absolute difference was consistently small across all models and synthetic datasets suggesting that these values could provide a reliable indicator of the expected decrease in accuracy in supervised machine learning models when developed using synthetic data. Health care departments could expect a manageable small yet consistent decrease in accuracy between real and synthetic data.

How Often Does the Winning Machine Learning Technique Change When Training Using Real Data to Training Using Synthetic Data?

The winning classifier when trained on synthetic data matched the winning classifier when trained on synthetic data in only 26% of cases for synthetic data generated using the CART and parametric methods and in just 21% of cases on data synthesized using the Bayesian network technique across the five machine learning models considered (SGD, DT, KNN, RF, and SVM). Tree-based methods were typically the winning classifier for models trained on real data; however, this was often not the case for models trained on synthetic data. When tree-based models were not considered, the winning classifier when trained on real data matched the winning classifier when trained on synthetic data in 74%, 53%, and 68% of cases for synthetic data generated using the CART, parametric, and Bayesian network approaches, respectively. It would appear that tree-based classifiers have some sensitivity to synthetic data, and the underlying cause requires further investigation.

What Is the Impact of Statistical Disclosure Control (ie, Privacy Protection) Measures on the Utility of Synthetic Data (ie, Similarity to Real Data)?

The average change in accuracy when SDC measures are applied to the training data was small across all machine learning models and for all SDC techniques. Again, tree-based models produced the largest decrease in accuracy across all SDC techniques. This is attributed to the synthetic data generation method and not the SDC measures, in line with previous results where SDC measures were not applied. We therefore conclude that the SDC techniques considered do not have a notable impact on data utility beyond what the synthetic data generation methods alone produce.

Limitations

This work has considered the impact of synthetic data on data utility when the data are used to train supervised machine learning algorithms. Further investigation with a broader range of machine learning algorithms, supervised and unsupervised, and including hyperparameter optimization is required. Such studies should cover an even larger range of datasets including, if possible, real health care department case studies.

Disclosure risk must also be explored in more detail. The impact of SDC measures on data utility has been considered in this work. Disclosure risk must also be measured across synthetic datasets, and a comparison of the data utility and disclosure risk trade-off should be performed.

Policy and Practice Implications

A wealth of rich health care data exists with the potential to provide new insights for the prevention of diseases, development of personalized medicine, and support of healthy life across the population. These data are held by health care data gatekeepers (eg, national health care departments) and are generally prevented from release, even for research purposes, due to justifiable privacy concerns around the protection of personal data, ethics, and in guaranteeing citizens' fundamental rights and freedoms.

Data sharing and data use demand careful governance, with legislation such as General Data Protection Regulation and the EU-US Privacy Shield placing increasingly stringent guidelines on data management. Data gatekeepers must manage myriad issues in relation to the nature of the data (eg, categories of sensitive data) and descriptions of the technical characteristics of processed data, as well as sharing and management of the data (eg, fair acquisition, data processing and data retention policies, legal basis for information processing, appropriate security measures) and the configuration of information systems that store and process the data.

From a health care perspective, a range of technical solutions using state-of-the-art machine learning could be developed using health care data with the potential to derive knowledge that can inform and enhance health care policy decision making and risk stratification [36,48]. Such tools can have a positive impact on health policy and practice, meeting the aims of national health departments, for example, as stated by the Department of Health Permanent Secretary in Northern Ireland, Richard Pengelly, in support of the MIDAS project, "the Department seeks to improve the health and social wellbeing of the people of NI, reduce health inequalities, and to assure the provision of appropriate health and social care services in clinical settings and in the community."

Accessing health care data to develop such tools is complex, involving a lengthy legal and ethical process, and in some cases access is impossible. Synthetic data can potentially overcome the barriers to accessing data and the need for compliance with data protection legislation as they infringe no privacy or confidentiality while remaining durable, reusable, shareable, clean, and potentially reliable as highlighted by Floridi [49], thus accelerating the development of machine learning to inform health care policy. Synthetic data also provide the opportunity to democratize the application of machine learning to health data for the benefit of patients and citizens enabling a larger community to leverage the power of machine learning in health care.

There is an increasing need for the development and evaluation of a robust and trustworthy synthetic data generator. Policy makers and clinicians who base decisions on models developed with synthetic datasets must be able to do so with the assurance that any knowledge elicited is very likely to be reflected in the real data. Using synthetic datasets to facilitate machine learning without disclosing sensitive data has the potential to revolutionize health care research and policy making in an impactful way by unlocking key research data in a secure way

that could drive improvements in population health and well-being much more quickly than is currently observed.

Conclusions

This work considers the efficacy of synthetic data for training supervised machine learning models for use by health care departments. The results are promising with small decreases in accuracy observed in models trained with synthetic data compared with those trained using real data. This work will be further extended to assist in the development of standard baselines for health care departments when using synthetic data (eg, an expected and acceptable decrease in accuracy) and synthetic data generators that can be trusted to produce the same winning model as that which would be produced by real data.

Acknowledgments

The MIDAS Consortium gratefully acknowledges the support to this project from the European Union research fund Big Data Supporting Public Health Policies under grant agreement No. 727721 (H2020-SC1-2016-CNECT SC1-PM-18-2016). We also acknowledge the support of the eHealth & Data Analytics Dementia Pathfinder Programme and Health and Social Care Board eHealth Directorate for this work under award ER/DARUG/09/18/10S.

Authors' Contributions

DR was mainly responsible for the paper. She designed and performed the experimental work and drafted the manuscript. All other authors gave feedback on various aspects of the paper including experimental design, machine learning, and policy and practice implications and revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Rumbold J, Pierscionek B. Contextual anonymization for secondary use of big data in biomedical research: proposal for an anonymization matrix. JMIR Med Inform 2018 Nov 22;6(4):e47 [FREE Full text] [doi: 10.2196/medinform.7096] [Medline: 30467101]
2. Luo G, Sward K. A roadmap for optimizing asthma care management via computational approaches. JMIR Med Inform 2017 Sep 26;5(3):e32 [FREE Full text] [doi: 10.2196/medinform.8076] [Medline: 28951380]
3. Dankar FK, Madathil N, Dankar SK, Boughorbel S. Privacy-preserving analysis of distributed biomedical data: designing efficient and secure multiparty computations using distributed statistical learning theory. JMIR Med Inform 2019 Apr 29;7(2):e12702 [FREE Full text] [doi: 10.2196/12702] [Medline: 31033449]
4. Suominen H, Zhou L, Hanlen L, Ferraro G. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. JMIR Med Inform 2015;3(2):e19 [FREE Full text] [doi: 10.2196/medinform.4321] [Medline: 25917752]
5. Reiter JP. New approaches to data dissemination: a glimpse into the future (?). CHANCE 2004;17(3):11-15. [doi: 10.1080/09332480.2004.10554907]
6. Rubin D. Statistical disclosure limitation. J Off Stat 1993;9(2):461-468. [doi: 10.1002/9781118445112.stat00072]
7. Little R. Statistical analysis of masked data. J Off Stat 1993;9:407-426.
8. Raghunathan T, Reiter J, Rubin D. Multiple imputation for statistical disclosure limitation. J Off Stat 2003;19:1-16.
9. Reiter J. Simultaneous use of multiple imputation for missing data and disclosure limitation. Survey Methodol 2004:235-242 [FREE Full text]
10. Reiter JP. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. J Royal Statistical Soc A 2005 Jan;168(1):185-205. [doi: 10.1111/j.1467-985x.2004.00343.x]
11. Reiter J. Significance tests for multi-component estimands from multiply imputed, synthetic microdata. J Stat Plan Infer 2005 May;131(2):365-377. [doi: 10.1016/j.jspi.2004.02.003]
12. Reiter J. Using multiple imputation to integrate and disseminate confidential microdata. Int Stat Rev 2009;77(2):179-195. [doi: 10.1111/j.1751-5823.2009.00083.x]
13. Reiter JP, Raghunathan TE. The multiple adaptations of multiple imputation. J Am Stat Assoc 2007 Dec;102(480):1462-1471. [doi: 10.1198/016214507000000932]
14. Reiter J, Drechsler J. Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. Stat Sinica 2007;20:405-422 [FREE Full text]
15. Reiter J. Using CART to generate partially synthetic public use microdata. J Off Stat 2005;21:441-462 [FREE Full text]
16. Ping H, Stoyanovich J, Howe B. DataSynthesizer: privacy-preserving synthetic datasets. 2017 Presented at: Proceedings of the 29th International Conference on Scientific and Statistical Database Management; 2017; Chicago. [doi: 10.1145/3085504.3091117]

17. Nowok B, Raab GM, Dibben C. synthpop: bespoke creation of synthetic data in R. J Stat Soft 2016;74(11). [doi: 10.18637/jss.v074.i11]
18. Heyburn R, Bond R, Black M, Mulvenna M, Wallace J, Rankin D, et al. Machine learning using synthetic and real data: similarity of evaluation metrics for different healthcare datasets and for different algorithms. 2018 Presented at: Proceedings of the 13th International FLINS Conference; 2018; Belfast. [doi: 10.1142/9789813273238_0160]
19. Dua D, Graff C. UCI machine learning repository. 2019. URL: http://archive.ics.uci.edu/ml [accessed 2020-06-27]
20. Dalianis H. Evaluation metrics and evaluation. In: Clinical Text Mining. Cham: Springer; 2018:45-53.
21. Domingo-Ferrer J, Torra V. Disclosure risk assessment in statistical data protection. J Comput Appl Math 2004 Mar;164-165:285-293. [doi: 10.1016/s0377-0427(03)00643-5]
22. Abowd J, Stinson M, Benedetto G. Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. 2006. URL: https://ecommons.cornell.edu/bitstream/handle/1813/43929/SSAfinal.pdf?sequence=3&isAllowed=y [accessed 2020-06-27]
23. Benedetto G, Stinson M, Abowd J. The Creation and Use of the SIPP Synthetic Beta. 2013. URL: https://ecommons.cornell.edu/bitstream/handle/1813/43924/SSBdescribe_nontechnical.pdf?sequence=3&isAllowed=y [accessed 2020-06-27]
24. Dushi I, Munnell AH, Sanzenbacher G, Webb A. Do households increase their savings when the kids leave home? SSRN Journal 2015. [doi: 10.2139/ssrn.2669704]
25. Chenevert R. Changing levels of spousal education and labor force supply. 2012. URL: https://www.census.gov/content/dam/Census/library/working-papers/2012/demo/SIPP-WP-263.pdf [accessed 2020-06-27]
26. Chung Y, Downs B, Sandler D, Sienkiewicz R. The parental gender earnings gap in the United States. 2017. URL: https://www2.census.gov/ces/wp/2017/CES-WP-17-68.pdf [accessed 2020-06-27]
27. Benedetto G, Gathright G, Stinson M. The earnings impact of graduating from college during a recession. 2010. URL: https://www2.vrdc.cornell.edu/news/wp-content/papercite-data/pdf/benedettogathrightstinson-11301.pdf [accessed 2020-06-27]
28. Carr MD, Wiemers EE. New evidence on earnings volatility in survey and administrative data. AEA Papers Proc 2018 May 01;108:287-291. [doi: 10.1257/pandp.20181050]
29. Greenstone M, Mas A, Nguyen H. Do credit market shocks affect the real economy? Quasi-experimental evidence from the great recession and "normal" economic times. Am Econ J Econ Policy 2020;12(1):200-225. [doi: 10.3386/w20704]
30. Kinney SK, Reiter J, Reznek A, Miranda J, Jarmin RS, Abowd JM. Towards unrestricted public use business microdata: the synthetic longitudinal business database. SSRN J 2011;79(3):362-384. [doi: 10.2139/ssrn.1759422]
31. Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L. Privacy: theory meets practice on the map. 2008 Presented at: Proceedings of the IEEE International Conference on Data Engineering; 2008; Cancun. [doi: 10.1109/icde.2008.4497436]
32. Hattersley L, Cresser R. Longitudinal Study 1971-1991: history, organisation and quality of data. 1995. URL: https://census.ukdataservice.ac.uk/media/51156/1971_defs.pdf [accessed 2020-06-27]
33. Boyle PJ, Feijten P, Feng Z, Hattersley L, Huang Z, Nolan J, et al. Cohort profile: the Scottish Longitudinal Study (SLS). Int J Epidemiol 2009 Apr;38(2):385-392. [doi: 10.1093/ije/dyn087] [Medline: 18492728]
34. O'Reilly D, Rosato M, Catney G, Johnston F, Brolly M. Cohort description: the Northern Ireland Longitudinal Study (NILS). Int J Epidemiol 2012 Jun;41(3):634-641. [doi: 10.1093/ije/dyq271] [Medline: 21296852]
35. Miranda J, Vilhuber L. Looking back on three years of Synthetic LBD Beta. 2014. URL: https://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1013&context=ldi [accessed 2020-06-27]
36. Black M, Rankin D, Wallace J, Bond R, Mulvenna M, Cleland B, et al. Meaningful integration of data, analytics and services of computer-based medical systems: the MIDAS touch. 2019 Presented at: Proceedings of IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); 2019; Cordoba. [doi: 10.1109/cbms.2019.00031]
37. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
38. Piai S, Claps M. Bigger data for better healthcare. 2013. URL: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/bigger-data-better-healthcare-idc-insights-white-paper.pdf [accessed 2020-06-27]
39. Rabesandratana T. European data law is impeding studies on diabetes and Alzheimer's, researchers warn. 2019. URL: https://www.sciencemag.org/news/2019/11/european-data-law-impeding-studies-diabetes-and-alzheimer-s-researchers-warn [accessed 2020-06-27]
40. Lugg-Widger FV, Angel L, Cannings-John R, Hood K, Hughes K, Moody G, et al. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: managing the morass. IJPDS 2018 Sep 12;3(3). [doi: 10.23889/ijpds.v3i3.432]
41. Lane J, Schur C. Balancing access to health data and privacy: a review of the issues and approaches for the future. Health Serv Res 2010 Oct;45(5 Pt 2):1456-1467 [FREE Full text] [doi: 10.1111/j.1475-6773.2010.01141.x] [Medline: 21054366]
42. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014 Jul;33(7):1123-1131. [doi: 10.1377/hlthaff.2014.0041] [Medline: 25006137]
43. MIDAS: meaningful integration of data, analytics and services. 2020. URL: http://www.midasproject.eu/ [accessed 2020-06-27]

44. Templ M, Meindl B, Kowarik A, Dupriez O. Simulation of synthetic complex data: the R package simpop. J Stat Soft 2017;79(10). [doi: 10.18637/jss.v079.i10]
45. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 2017 Aug 30;25(3):230-238. [doi: 10.1093/jamia/ocx079] [Medline: 29025144]
46. Elliot M. Final report on the disclosure risk associated with the synthetic data produced by the SYLLS team. 2015. URL: https://www.cmi.manchester.ac.uk/research/publications/reports/ [accessed 2020-06-27]
47. Ritchie F, Elliot M. Principles- versus rules-based output statistical disclosure control in remote access environments. IASSIST Q 2015 Dec 11;39(2):5. [doi: 10.29173/iq778]
48. Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, et al. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. JMIR Med Inform 2020 Feb 20;8(2):e16492 [FREE Full text] [doi: 10.2196/16492] [Medline: 32130148]
49. Floridi L. What the near future of artificial intelligence could be. Philos Technol 2019 Mar 19;32(1):1-15. [doi: 10.1007/s13347-019-00345-y]

Abbreviations

AI: artificial intelligence
CART: classification and regression tree
DT: decision tree
KNN: k-nearest neighbors
MIDAS: Meaningful Integration of Data, Analytics, and Services
RF: random forest
SDC: statistical disclosure control
SGD: stochastic gradient descent
SVM: support vector machine

Edited by G Eysenbach; submitted 26.03.20; peer-reviewed by F Dankar; comments to author 20.04.20; revised version received 24.04.20; accepted 04.06.20; published 20.07.20.

Please cite as:
Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G
Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing
JMIR Med Inform 2020;8(7):e18910
URL: http://medinform.jmir.org/2020/7/e18910/
doi:10.2196/18910
PMID:32501278

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org