JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 10 (2022), Issue 8 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Viewpoints

Original Papers

Corrigenda and Addendas

Review

Application of Artificial Intelligence in Shared Decision Making: Scoping Review

Samira Abbasgholizadeh Rahimi[1,2,3], BEng, PhD; Michelle Cwintal[4], MSc; Yuhui Huang[5], MSc; Pooria Ghadiri[1], MD; Roland Grad[1], MSc, MD; Dan Poenaru[6], MHPE, MD, PhD; Genevieve Gore[7], BA, MSc; Hervé Tchala Vignon Zomahoun[8,9], MSc, PhD; France Légaré[8,9,10], MD, PhD; Pierre Pluye[1], MD, PhD

[1]Department of Family Medicine, McGill University, Montreal, QC, Canada

[2]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada

[3]Mila-Quebec AI Institute, Montreal, QC, Canada

[4]Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, QC, Canada

[5]Department of Integrated Studies in Education, McGill University, Montreal, QC, Canada

[6]Department of Pediatric Surgery, McGill University Health Centre, Montreal, QC, Canada

[7]Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montreal, QC, Canada

[8]Centre de recherche en santé durable, Centre intégré universitaire de santé et services sociaux de la Capitale-Nationale, Quebec City, QC, Canada

[9]Quebec Support for People and Patient-Oriented Research and Trials Unit, Quebec City, QC, Canada

[10]Department of Family Medicine and Emergency Medicine, Faculty of Medicine, Université Laval, Quebec City, QC, Canada

Corresponding Author:
Samira Abbasgholizadeh Rahimi, BEng, PhD
Department of Family Medicine
McGill University
5858 Cote-des-Neiges Rd, Suite 300
Montreal, QC, H3S 1Z1
Canada
Phone: 1 (514)399 9218
Email: samira.rahimi@mcgill.ca

Abstract

Background: Artificial intelligence (AI) has shown promising results in various fields of medicine. It has the potential to facilitate shared decision making (SDM). However, there is no comprehensive mapping of how AI may be used for SDM.

Objective: We aimed to identify and evaluate published studies that have tested or implemented AI to facilitate SDM.

Methods: We performed a scoping review informed by the methodological framework proposed by Levac et al, modifications to the original Arksey and O'Malley framework of a scoping review, and the Joanna Briggs Institute scoping review framework. We reported our results based on the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) reporting guideline. At the identification stage, an information specialist performed a comprehensive search of 6 electronic databases from their inception to May 2021. The inclusion criteria were: all populations; all AI interventions that were used to facilitate SDM, and if the AI intervention was not used for the decision-making point in SDM, it was excluded; any outcome related to patients, health care providers, or health care systems; studies in any health care setting, only studies published in the English language, and all study types. Overall, 2 reviewers independently performed the study selection process and extracted data. Any disagreements were resolved by a third reviewer. A descriptive analysis was performed.

Results: The search process yielded 1445 records. After removing duplicates, 894 documents were screened, and 6 peer-reviewed publications met our inclusion criteria. Overall, 2 of them were conducted in North America, 2 in Europe, 1 in Australia, and 1 in Asia. Most articles were published after 2017. Overall, 3 articles focused on primary care, and 3 articles focused on secondary care. All studies used machine learning methods. Moreover, 3 articles included health care providers in the validation stage of the AI intervention, and 1 article included both health care providers and patients in clinical validation, but none of the articles included health care providers or patients in the design and development of the AI intervention. All used AI to support SDM by providing clinical recommendations or predictions.

Conclusions: Evidence of the use of AI in SDM is in its infancy. We found AI supporting SDM in similar ways across the included articles. We observed a lack of emphasis on patients' values and preferences, as well as poor reporting of AI interventions,

resulting in a lack of clarity about different aspects. Little effort was made to address the topics of explainability of AI interventions and to include end-users in the design and development of the interventions. Further efforts are required to strengthen and standardize the use of AI in different steps of SDM and to evaluate its impact on various decisions, populations, and settings.

(JMIR Med Inform 2022;10(8):e36199) doi:10.2196/36199

KEYWORDS

artificial intelligence; machine learning; shared decision making; patient-centered care; scoping review

Introduction

Shared Decision Making

Shared decision making (SDM) is the process in which patients and health care providers collaborate to make decisions based on the latest medical evidence and patients' preferences and values [1]. In this process, health care providers present the patient with options for screening or treatment and evidence for each option's harms and benefits. The patient is invited and supported in expressing their preferences and values, and eventually, patients and their health care providers collaboratively make a decision that is best aligned with patients' preferences and values [1]. Thus, the final shared decision is informed by the best evidence and by what matters most to the patient [2]. The use of SDM in clinical practice has been limited [3-5]. The most frequently reported reasons by health care providers are time pressure, lack of applicability because of patient characteristics, and clinical situations [6].

Elwyn et al [7,8] presented a 3-step model for clinical practice, consisting of team talk, option talk, and decision talk. During team talk, the need to provide support to patients when choices are presented and to elicit their goals to guide decision-making is emphasized. Option talk consists of providing more information about these options and comparing them through risk communication. Finally, during decision talk, health care providers guide patients to a decision based on their experience and expertise, which reflects the informed preferences of patients. The model aims to simplify the process so that health care providers can integrate SDM and patient decision support into their practice. Despite this, the use of SDM in clinical practice faces barriers that can potentially be alleviated by using artificial intelligence (AI).

Artificial Intelligence and Its Potential in Health Care

AI, defined as "computational intelligence" or the "science and engineering of making intelligent machines" [9], describes the fast-growing field of simulating intelligent, human-like behavior in computers and technology [10]. AI can provide decisional support to health care providers and patients. Machine learning, a subfield of AI, enables computers to learn from data without explicit programming [11,12]. Computers are provided with large data sets and learn to make accurate predictions, for example, on the diagnosis and prognosis of health outcomes of an individual, in different settings, including primary health care [13], identifying patterns and trends and learning from previous experience [14].

In the last 2 decades, AI has been applied in various fields, such as telecommunications [15], financial services [16], and health care [17]. AI has shown promising results in various fields, including radiology [18], ophthalmology [19], cardiology [20], orthopedics [21], and pathology [22]. For example, in medical imaging, AI can be used to assess x-rays, thus reducing the workload of health care providers [23]. It also has the potential to help health care providers assess patients' health risks, increase the efficiency and effectiveness of intervention and treatment, empower patients to better understand their health and self-manage their conditions, reduce waiting times and costs, and ultimately improve the quality of care and patient outcomes [24-26].

AI has the potential to foster SDM by informing decision-making and allowing health care providers to focus their energy on spending more time with the patient [27]. AI tools provide a wide variety of information with the ability to analyze large amounts of data and discover correlations that may have been missed by researchers and health care providers [28]. There is emerging literature regarding the bioethics and obstacles behind using AI for health decision-making [27], patients' and health care providers' perceptions of AI-based decision aids [29] and how it should be incorporated to ensure that health care is patient-centered. However, little is known about how AI is used in SDM in practice and how it can facilitate the decision-making step of SDM. Therefore, we aimed to systematically examine the evidence on the use of AI in SDM through a scoping review to map existing knowledge.

Objective and Research Question

The objective of the scoping review is to examine evidence on the use of AI in SDM, namely, to explore what has already been done and what future roles may exist for the use of AI in SDM.

Our specific research questions are as follows: (1) What is the available knowledge on the use of AI interventions for SDM? (2) How is AI being used for the decision-making point of SDM?

Methods

Study Design

The scoping review methodological framework proposed by Levac et al [30], modifications to the original framework of a scoping review [31], and the Joanna Briggs Institute methodological guidance for scoping reviews [32] were used to guide this research. We developed a protocol with the following steps: (1) identifying the research question; (2) identifying relevant studies; (3) selecting studies using an iterative team approach to study selection and data extraction; (4) charting the data by incorporating a numerical summary; (5) collating, summarizing, and reporting the results; and (6) consulting the results regularly. This protocol is registered and

available on the Open Science Framework website [33]. We completed this review according to the published protocol. Finally, we used the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist for reporting [34]. The filled PRISMA-ScR checklist is available in Multimedia Appendix 1.

Eligibility Criteria

We defined the eligibility criteria for our search using the *Population, Intervention, Comparison, Outcomes, Setting and study designs* components [35].

Population

Any population that provided health care (eg, general practitioners, nurses, social workers, pharmacists, and public health practitioners) and any individual who received care (eg, patients and their families and caregivers) were included.

Intervention

Any AI intervention implemented or tested during an SDM process in a clinical context was included in the study. AI was defined according to the definition provided by McCarthy [9] and Russell et al [36]. AI interventions included expert systems, knowledge representation, machine learning involving predictive models, reinforcement learning, natural language processing, and computer vision. If the AI intervention was not used for the decision-making point in SDM, it was excluded. We defined SDM as a process that occurred if the following three steps had taken place: (1) team talk, (2) option talk, and (3) decision talk [7,8].

Comparators or Control

No limitation.

Outcome

Any outcome related to patients, health care providers, or health care systems were included in this study.

Setting and Study Design

Studies in any health care setting (eg, primary care and secondary care); all studies using qualitative, quantitative, and mixed methods designs; and only studies published in the English language were included. Reviews, opinion pieces, editorials, comments, news articles, letters, and conference abstracts were excluded.

Information Sources and Search Strategy

A comprehensive literature search was designed and conducted by an experienced information specialist in consultation with the research team. The seed articles were identified by experts on the team, and the final search strategy was reviewed by the lead author. The process of the literature search was iterative. The following six electronic databases were searched from their inception to May 2021: MEDLINE (Ovid), EMBASE (Ovid), Web of Science Core Collection, CINAHL, Cochrane Library (CENTRAL), and IEEE Xplore Digital Library. The reference lists of the included studies were searched manually. Retrieved records were managed with EndNote X9.2 (Clarivate) and imported into the DistillerSR review software (Evidence Partners) to facilitate the selection process. The final search strategies and key terms for each database are available upon request.

Study Selection Process

We removed duplicates and then applied the inclusion criteria for level 1 (title and abstract) and level 2 (full text) screening using a standardized inclusion criteria grid. A pilot test of 55 studies (12% of the total 458 citations) for level 1 screening was conducted. Once familiar with the literature of interest, we modified the a priori eligibility criteria to adjust our study selection where necessary. Subsequently, 2 reviewers (PG, MC, and YH) independently screened the titles and abstracts. The reasons for exclusion were recorded for full-text selection. Any disagreements regarding study inclusion were resolved by a third reviewer (SAR).

Data Items and Data Collection Process

A data extraction form was drafted and finalized with feedback from the team members. Elements for data extraction included study characteristics (eg, year published, country of the corresponding author, and study setting), characteristics of the AI intervention (eg, purpose of the intervention, methods/techniques used, data sources, and performance), involvement of end users in the development of the intervention (eg, health care providers and patients), aspects of the AI intervention (eg, explainability of AI and reproducibility of intervention), whether AI was implemented or tested, how the AI intervention was used for decision-making in SDM, and outcomes (eg, related to patients, health care providers, and health care systems). A total of 2 reviewers (YH, PG, and MC) independently extracted relevant data from each included study. All data were verified by a third reviewer (SAR).

Critical Appraisal

In alignment with the proposed framework for methodological guidance in scoping reviews, we did not conduct a quality appraisal. Critical appraisal in scoping reviews is not considered mandatory [30-32].

Synthesis

We summarized our findings using descriptive statistics and performed a narrative synthesis describing the characteristics of the AI intervention, whether end users were involved in the development and/or its validation, how the AI intervention supported the decision point of SDM, and what the outcomes were if it was implemented in a clinical setting. We informed our synthesis through the work and toolkits published by Popay et al [37], titled “Guidance on the conduct of narrative synthesis in systematic reviews.” Specifically, we performed a thematic analysis and identified 3 main themes across the included studies in an inductive manner (involvement of end users, outcomes of AI interventions, and AI interventions for the decision point). This allowed us to organize and present our results comprehensively.

Consultation

The results were provided to the team members for their feedback. Study updates were also provided to the researchers and health care providers during 2 workshops led by the first

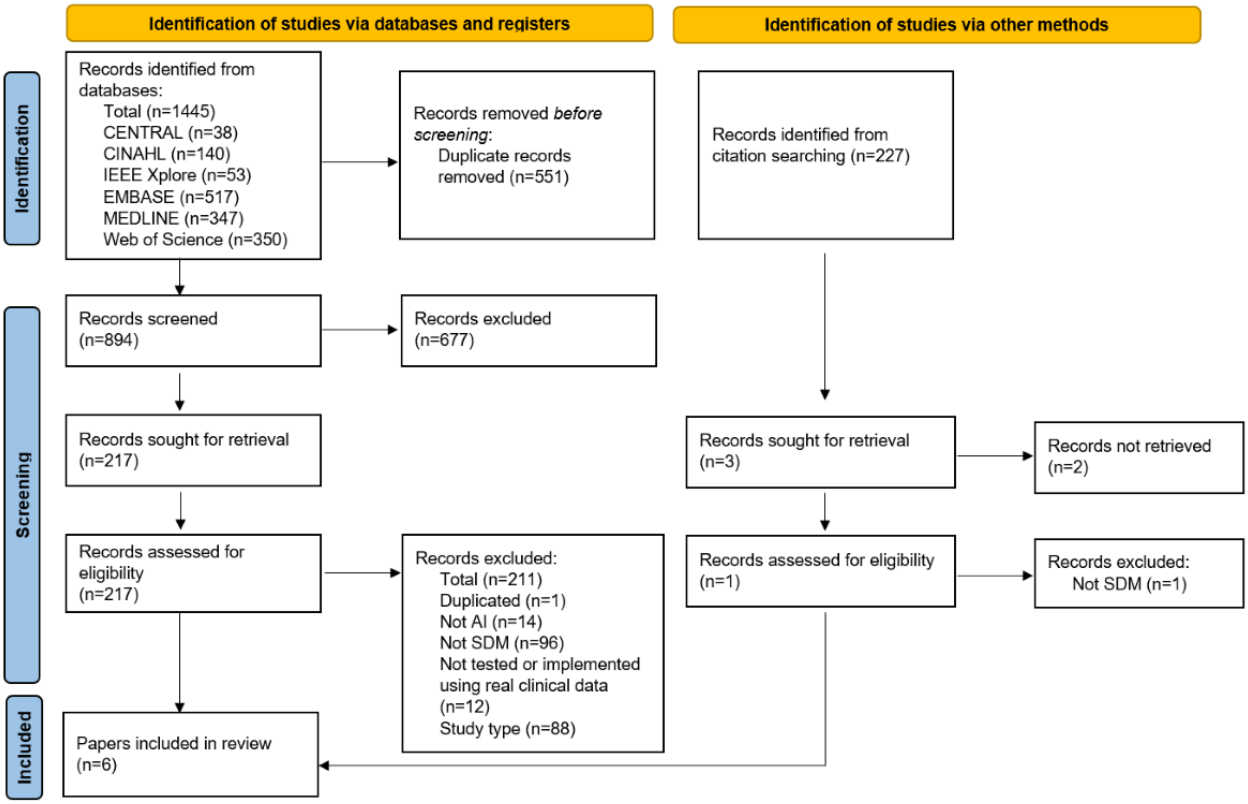
author (SAR) at 2 international scientific conferences, that is, the 10th International Shared Decision Making Conference and the annual meeting of the North American Primary Care Research Group.

Results

Flow of Studies

The search process resulted in 1445 records from the selected electronic databases, 551 of which were excluded as duplicates. Of the remaining 894 studies, we excluded 677 at level 1 screening because they did not meet the inclusion criteria and the remaining 217 underwent full-text review. Citations were manually searched (n=227), of which 3 studies were sought for retrieval and was assessed for eligibility. No eligible studies were found in the reference search. Ultimately, 6 articles met our inclusion criteria (Figure 1). Of 6 articles, 2 referred to the same study [38,39]. The full list of included articles and their details can be found in Multimedia Appendix 2 [34-39].

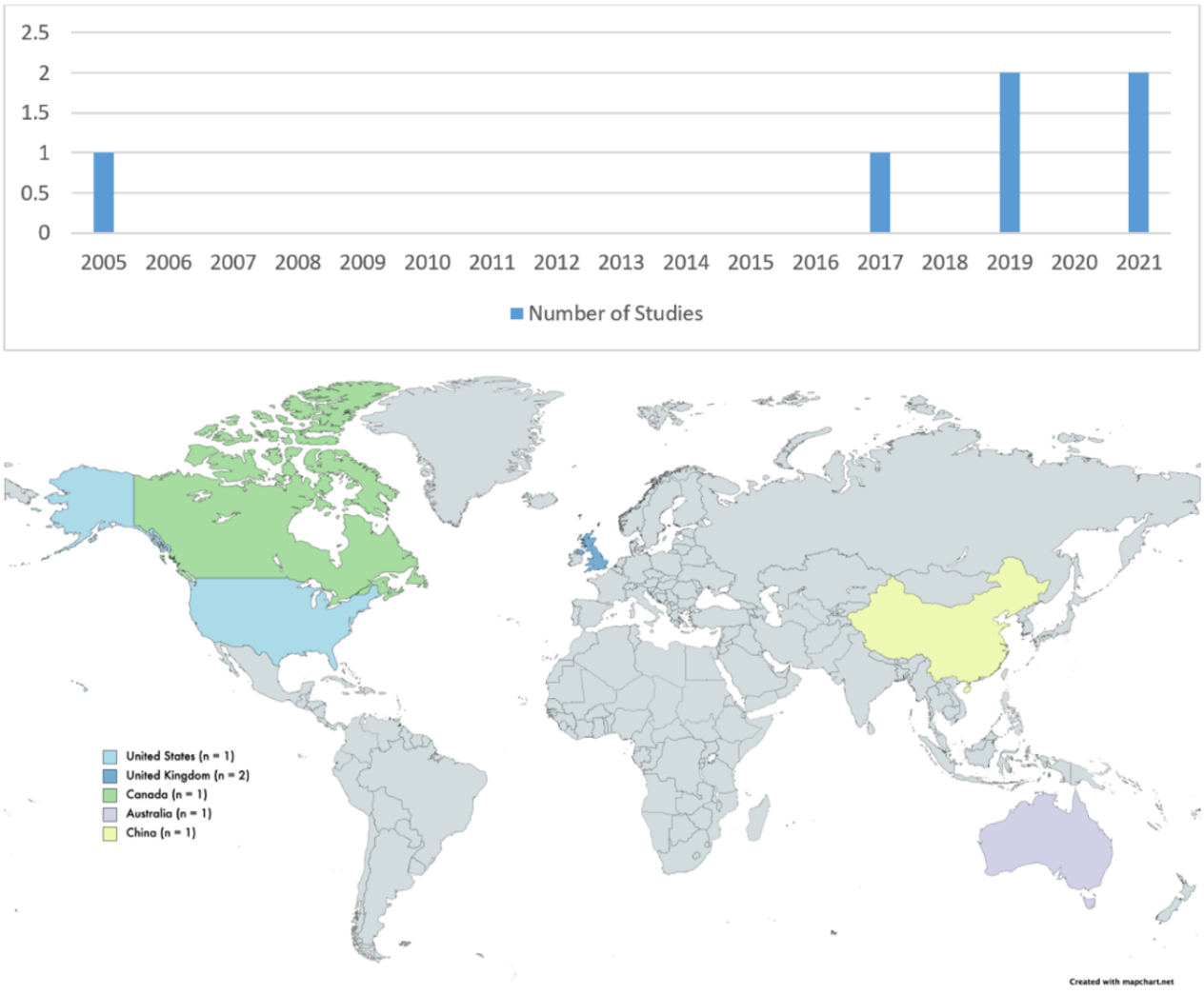
Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. Adapted from Page et al [40]. AI: artificial intelligence; SDM: shared decision making.



Characteristics of Included Articles

The number of studies published annually has increased since 2017, with the majority conducted in North America and Europe. The distribution and publication dates of the included studies are shown in Figure 2.

Figure 2. Years of publication and countries where studies are outlined in the included papers.



AI Characteristics—Purpose, Development, Data Sets, and Performance

In Table 1, we highlight the AI characteristics of the included studies, such as the AI method used, characteristics of the data set, and performance measures.

Table 1. Characteristics of artificial intelligence (AI) interventions.

Study	AI method	Data set and its characteristics	Performance
Frize et al [41]	Machine learning, artificial neural networks, and case-based reasoning	• Not provided	Not provided
Wang et al [42]	Machine learning, multilabel classification methods, k-nearest neighbors, and random k-label sets	• Electronic health records • 2542 patients • 65.6% male, 34.4% female • Mean age 66.46 (SD 13.81) years • 70% of this was used for training, and 30% was used for testing	Performance accuracy of 0.76
Twiggs et al [43]	Machine learning, Bayesian belief network, and Bayes network	• Data from the National Institutes of Health Osteoarthritis Initiative • 330 patients, between the ages of 45 and 79 years, have undergone total knee arthroplasty	Not provided
Jayakumar et al [44]	Machine learning (type not specified)	• Not provided	Not provided
Kökciyan et al[a] [38,39]	Metalevel argumentation frameworks	• Not provided	Not provided

[a]This refers to both articles describing the system developed by Kökciyan et al [38,39], which were included.

Of the included articles, all used machine learning as the type of AI. Only 2 articles presented information on the data set used to develop the AI intervention [42,43], and 1 article reported the performance accuracy (0.76) of their intervention [42].

Most of the included articles (n=4) did not report on the data set used to develop the AI intervention; among those that did (n=2), only 1 reported on the sex distribution of the patient data [42], and both provided information about the age (mean or range) of patients in their data set. Only 1 article specifically mentioned the breakdown of data used to develop and test their intervention [42] but did not report data set characteristics for the 2 breakdowns. None of the included articles commented on the generalizability of the algorithm or representativeness of the data used to develop and train the AI intervention. Although 2 articles mentioned the aspects of explainability and interpretability [39,43], none of the included articles reported on how they developed their AI interventions to be explainable and/or interpretable.

Explainable AI is a broad and new domain and is being studied in AI. In general, we can consider explainability throughout AI development: (1) *premodeling explainability*, (2) *explainable modeling*, and (3) *postmodeling explainability*. One of the challenges in this field is the so-called explainability versus performance trade-off (often, high-performance methods such as deep learning are less explainable).

In health care, explainability and interpretability are required for patients and health care providers to understand why AI interventions produce a certain prediction or suggestion and to trust this output [45]. Without this understanding, ethical and practical challenges arise, including a lack of trust and transparency in AI tools [28]. A lack of explainability and interpretability creates an informational discrepancy between patients and health care providers, impeding risk assessment and giving rise to ethical issues such as the ascription of responsibility when an incorrect decision is made [28]. Moreover, a lack of explainability and interpretability ties into the issue of informed consent in health care [46]. It is unclear as to what level of understanding patients who use AI require to provide informed consent and to what extent health care providers are responsible for educating their patients on its use [46]. However, explainability and interpretability are crucial in increasing the transparency of the inner workings of the system and in fostering the trust of health care providers and patients in the outcomes the AI may provide throughout the process of SDM [45].

Frize et al [41] developed and tested a decision support system that used AI to tailor information to help parents decide to continue, limit, or discontinue intensive care of newborns [41]. Machine learning methods, such as artificial neural networks and case-based reasoning methods, were used in this decision support system. The AI component was capable of knowledge learning, processing, and derivation. The developed system was able to provide structuralized knowledge translation and exchange between all participants and facilitate collaborative decision-making. Overall, clinicians found the classification rates of the model acceptable in comparison with the constant predictor used as a statistical benchmark, but no other performance metrics were presented.

Wang et al [42] proposed an SDM system framework connected to the electronic health records (EHRs) of patients with type 2 diabetes to provide them and their health care providers with tailored knowledge and choices about medications [42]. Machine learning methods, multilabel classification methods including k-nearest neighbors algorithms, and random k-label sets using EHR data were used to provide a medication recommendation list based on patients' current conditions. The data set used to develop the AI intervention included data from 2542 patients. Of these, 65.62% (1668/2542) were men and 34.38% (874/2542) were women. The mean age of the included patients was 66.46 (SD 13.81) years. Associated diseases and vital sign values were also reported. The authors used 70% of the total data set to train

the AI algorithm, and the remaining 30% to test it. The AI model had an accuracy of 0.76.

Twiggs et al [43] developed a clinical tool to predict total knee arthroplasty outcomes for patients with advanced osteoarthritis to help patients and surgeons decide whether a surgical or nonsurgical pathway is most appropriate on a patient-specific basis. The group developed a Bayesian belief network to identify patients at risk of limited improvement from total knee arthroplasty using data from the National Institutes of Health Osteoarthritis Initiative, a publicly accessible database. A total number of 330 patients between the ages of 45 and 79 years who had undergone total knee arthroplasty were included. The team used a machine learning method, that is, a naive Bayes network, for variable selection and model generation.

Jayakumar et al [44] performed a randomized clinical trial to assess whether an AI-based decision aid influenced decision quality, patient experience, functional outcomes, and process-level outcomes in patients with advanced osteoarthritis considering total knee replacement. They used a machine learning–based platform to generate personalized outcomes. Neither the development nor the performance of the model was described in the article; however, they mentioned that the AI intervention had been tested in a clinical setting and that its fidelity had been discussed with the clinical team before deployment.

Kökciyan et al [38,39] developed a decision support system, that is, "CONSULT," to help patients who had stroke in self-management and adherence to treatment plans, in collaboration with health care providers. Patients, caregivers, and health care providers collaborate to decide the best treatment plan for the patient. The system was developed using metalevel argumentation frameworks. Wellness sensor data, EHR data, and clinical guidelines were used as input, and recommendations and textual explanations for automated decisions were provided as output.

Involvement of the End Users

In terms of end user (ie, patients and health care providers) involvement in the design, development, and/or validation of AI systems, we found that 3 of the articles [39,41,44] included health care providers to validate the AI intervention, and 1 of the articles included both health care providers and patients in clinical validation of their AI tool [43]. The first 3 articles involved clinicians validating the correctness of the recommendations and explanations provided to patients by the AI tool [39], confirmed the fidelity of the AI intervention before deployment [44], and were included in the testing of usability and acceptability as well as a needs assessment of the intervention [41]. Twiggs et al [43] clinically validated their tool for both patients and health care providers.

One of the articles [38] also held initial patient focus groups in which co-design activities were held. These activities resulted in a user-centered version of how they wished to see the information displayed by the decision support tool. No additional information on how the co-design activities were organized was provided.

Population Characteristics and Outcomes

In total, 4 of the included articles tested their interventions for usability and acceptability [38,39,41,42], and 2 of the articles implemented their interventions in clinical settings with targeted end users (eg, patients and health care providers) [43,44]. Only the last 2 articles reported outcomes related to patients and health care providers. These were primarily psychosocial outcomes and included better decisional quality, improved SDM, increased satisfaction, and better clinical postoperative outcomes. Of the included articles, 3 also reported outcomes related to health care systems [42-44]. These were related to the general workflow and how the interventions did not significantly alter the flow or time it took to provide care. They also include the high feasibility and convenience of integrating AI into health care systems.

All the included articles provided some level of detail related to the population of the data sets that they used to train or test their algorithm. Only 1 article provided a thorough presentation of the population by reporting the sociodemographic characteristics of the participants involved [44]. In total, 4 articles tested the interventions for usability and acceptability, whereas 2 articles observed actual outcomes by applying their intervention in clinical contexts [43,44].

Frize et al [41] tested their AI for acceptability and usability with an expert panel consisting of a neonatologist, engineer or computer scientist, clinical nurse specialist, social worker, and ethicist. The classification rate of the intervention was found to be acceptable for a clinical trial tool. The needs assessment performed through interviews with 5 neonatal clinicians confirmed that the design of their tool met the needs of the population for which it was designed. Acceptability was evaluated using open-ended questions based on a questionnaire from the Foundation for Informed Medical Decision-Making. The expert panel found the tool clear and easy to use.

Kökciyan et al [38,39] performed a pilot study using their CONSULT system to assess its usability and acceptability. The system was implemented as a mobile Android app, and 6 healthy volunteers were recruited to use the system for a week. They interacted with different aspects of the system and were asked to regularly collect measurements from wellness sensors and input data. A pilot study demonstrated the usability of the app.

Wang et al [42] tested their AI interventions using clinical data. The authors used 30% of the clinical data set mentioned earlier to test the AI intervention. The total data set included data on 2542 patients, of which 65.6% (n=1668) were male. As these EHRs only included hospitalized patients, the outcome of medication use was not considered. In terms of outcomes for health care systems, the intervention was reported to have high feasibility and maintenance—if the model or knowledge required for proper function became outdated, the intervention could be modified without affecting the normal operation of the hospital's EHR system.

Jayakumar et al [44] conducted a randomized clinical trial that recruited 129 patients with presumptive knee osteoarthritis who were candidates for primary total knee replacement. A total of 69 patients were in the intervention group (n=46, 67% women)

and 60 were in the control group (n=37, 62% women). The mean age of the intervention group was 62.59 (SD 8.85) years, whereas the mean age of the control group was 62.62 (SD 7.81) years. The authors reported on ethnicity, education, work status, social status, and insurance status for both the intervention and control groups. The control group received an educational module and usual care, whereas the intervention group received a preference model and an output from the AI tool. Both groups met the surgeons afterward for the decision-making discussion. In terms of patient-related outcomes, the intervention group showed better decisional quality and improved SDM, patient satisfaction, and functional outcomes. Overall, the use of the AI tool did not prolong consultation times.

Twiggs et al [43] performed a clinical validation with 150 patients who presented to a surgeon with >30 years of experience in 2 cohorts. They included patients aged ≥55 years with knee pain without a history of meniscal or ligamentous injury. They collected data over 3 months. Patients were first asked to fill a digital questionnaire based on knee osteoarthritis and injury outcome scores, as well as demographic and medical condition data. These data were used by their developed intervention to calculate a predictive postoperative score and display it visually on a percentile scale of the pain of a population of patients with osteoarthritis seeing a surgeon. The first cohort consisted of 75 (50%) consenting patients who filled the group's developed questionnaire. In this cohort, the surgeon and patients were blinded to the predictive output of the tool and proceeded with their consultations as normal. The second cohort consisted of 75 (50%) consenting patients, and both the patients and surgeons were exposed to the output of the intervention. The outcomes were reported for patients and surgeons. Although the use of the AI intervention output did not change the proportion of patients booked for total knee arthroplasty surgery, there was a change in the level of patient-reported pain between those booked and not booked for surgery when using the tool. Apart from the questionnaire, which only took 10 minutes to complete, there was no disruption to the normal surgeon consultation workflow.

AI Interventions for the Decision Point

Of the included articles, 3 designed AI interventions for primary care [38,39,42], relating to the care of individuals with chronic conditions including patients with diabetes and stroke survivors, and 3 for secondary care [41,43,44], of which 2 (67%) focused on patients requiring treatment for their knee and 1 (33%) focused on neonatal intensive care. The included articles supported the decision-making step of SDM by introducing interventions to predict outcomes [41,43,44] of clinical significance and for clinical recommendations [38,39,42]. In Table 2, we provided information about the setting, decision-making problem, and a summary of how AI is being used for decision-making in SDM.

Table 2. Summary of artificial intelligence interventions and how they are being used for decision-making in the included studies.

Study	Setting	Decision-making problem	AI[a] for decision-making
Wang et al [42]	Primary care	Knowledge and choices about antihyperglycemic medications	The tool provides patients and health care providers with tailored knowledge and choices about antihyperglycemic medications through the integration of electronic health record data. Patients and physicians can review patients' conditions more comprehensively and tailor consultations to the patient's current condition.
Frize et al [41]	Secondary care	Neonatal intensive care decisions	The tool allows health care providers to predict outcomes in neonatal intensive care and counsel families on the pros and cons of deciding to initiate or withdraw treatment. The tool also promotes parental involvement in the decision-making process.
Twiggs et al [43]	Secondary care	The decision about total knee arthroplasty	The AI intervention presents end users (patients and surgeons) with interpretable information relating to the risk of no improvement after total knee arthroplasty. This helps them decide whether to proceed with total knee arthroplasty.
Jayakumar et al [44]	Secondary care	The decision about total knee replacement	AI system provides patients with a personalized outcome report, which is then discussed with the surgeon during decision-making discussions.
Kökciyan et al [38,39][b]	Primary care	The decision about treatment plans and options for stroke survivors	This tool supports the decision-making point by providing an up-to-date view of the patients' situation based on personalized metrics and provides explanations for its recommendations.

[a]AI: artificial intelligence.

[b]This refers to both articles describing the system developed by Kökciyan et al [38,39] that were included.

The AI intervention by Wang et al [42] supports the decision point by providing patients and health care providers with tailored knowledge and choices about antihyperglycemic medications through the integration of EHR data. Their tool was designed with specific end-user interfaces for each step of SDM (team talk, option talk, and decision talk). During decision

talk, patients can have more efficient conversations with their health care providers based on the medication recommendations that the AI system provides. It is designed for both inpatient and outpatient settings and provides a more intuitive understanding of patient conditions and knowledge of diabetes medications.

The AI intervention by Frize et al [41] supports the decision point as the components of the tool interact to provide predictive analysis, document repository, customized delivery, and adaptive interfaces. They aimed to augment group clinical processes in various phases of decision-making. The goal was to promote parental involvement and collaboration with the clinical team. The tool allows health care providers to predict outcomes in neonatal intensive care and counsel families on the pros and cons of deciding to initiate or withdraw treatment.

The tool presented by Twiggs et al [43] supports the decision point by presenting end users, that is, patients and surgeons, with interpretable information relating to the risk of no improvement following total knee arthroplasty. It provides interpretable output, allowing end users to understand the impact of alternative treatments. This tool helps patients and their surgeons decide whether they are good candidates for the procedure.

The intervention by Jayakumar et al [44] supports the decision point by providing patients with a personalized outcome report based on data inputs (ie, demographics, patient-reported outcome measurements, and clinical comorbidities), which is discussed with the surgeon during the decision-making.

The CONSULT system by Kökciyan et al [38,39] supports the decision-making point in SDM by presenting an up-to-date view of the patient's situation based on personalized metrics, from a patient's EHR and wireless sensor input and providing textual explanations of automated decisions of the tool to accompany the recommendations it provides. The relevant, up-to-date, summarized data CONSULT provides, along with treatments and recommendations, support the decision-making point between patients and their health care professionals.

Discussion

Principal Findings

We conducted a scoping review as a first step toward a comprehensive overview of the literature on the use of AI in SDM. This overview provides a basis for future systematic review. The results of our study lead us to make the following observations.

Role of AI in SDM

The included articles presented AI interventions used for decision-making during SDM in similar ways. Within the included articles, AI interventions were specifically applied to predict outcomes of clinical significance and for clinical recommendations. The decision-making step can benefit from AI interventions because AI can present a comprehensive and personalized list of treatment options, as well as risks and benefits, thus increasing the amount of knowledge related to the condition, treatment, side effects, risks, and outcomes. AI models are capable of learning and processing all information related to a patient's care and can generate evidence-based recommendations to support SDM [47]. These models can also be used to support risk communication. Similar to how they may be integrated into an intelligent tutoring system, predictive models can present relevant information when discussing risks associated with a patient's condition in a manner appropriate for that specific patient, as well as assess their level of understanding and provide supplementary information accordingly [48].

The decision-making step is a core step of SDM, in which patient–health care provider interaction is essential and should remain independent of and unrestrained by AI intervention. Patient–health care provider relationships are based on responsibilities that provide a foundation for the relationship to grow. Despite acknowledging the benefits AI may have on facilitating SDM, patients continue to expect their health care provider to retain final discretion over treatment plans and monitor their care, as well as to adapt any contribution from the AI intervention to their unique situation [49]. Conversely, patients expect to remain empowered in decision-making and can either dispute or refuse the input of AI [49]. It is important to design and implement AI interventions in clinical settings in a way that does not negatively impact the human and personal aspects of certain decisions during the SDM process. AI interventions must be implemented in ways that preserve and uplift patient–health care provider relationships in care, as well as facilitate making shared medical decisions.

AI interventions can open up more time for health care providers to spend connecting with their patients; however, they may place the health care provider in a mediator-like role, in which they will be responsible for explaining the AI output to their patients. This can be difficult to achieve, especially when a lack of interpretability and explainability may exist in certain AI models, such as deep learning. This lack of interpretability and explainability can result in a lack of trust and decisional delay or conflict consequently, which are factors that SDM aims to resolve [27]. AI interventions in health care can influence patient–health care provider relationships [27], but little is known about how they influence this relationship and what are the best ways to integrate AI into SDM, to use its benefits and mitigate potential risks. Further work is required to investigate how the different steps of SDM can benefit from AI intervention without affecting the patient–health care provider relationship.

Explainability and Interpretability of AI Systems

One of the principal challenges in the incorporation of modern AI interventions into health care is explainability and interpretability. This refers to the insight an AI intervention gives to clarify its function to an audience; that is, *how* an algorithm generates output from a given input [50-52]. The levels of explainability and interpretability depend on the AI method used. This is the case in certain AI models such as deep learning.

Despite the promising performance of AI, its implementation in clinical practice remains challenging. Trust in AI is one of the main barriers to its adoption in clinical practice [53]. The inability of humans to understand why an AI system makes

particular decisions limits the effectiveness of the new generation of AI systems in critical settings, such as primary health care. Prior work has highlighted the significance of explainable AI in health care and has shown that the lack of explainability (*black box*) in AI systems can affect physicians' and patients' trust in AI [54-56].

In our review, 2 of the included articles [39,43] briefly touched on explainability and interpretability, stating that textual explanations were provided by the AI tool to explain automated decisions [39] and that the outcome of their AI model is interpretable [43]. However, these 2 articles did not explain the steps they had taken in the development of their tool to make it explainable or interpretable, and none of the other included articles considered these aspects. This might introduce barriers to the implementation of these systems in the process of SDM in clinical practice. As in any other context that attempts to integrate AI into sensitive human interactions, AI explainability, and interpretability for SDM needs to be addressed.

Moreover, the level of understanding of the explainability and interpretability of AI tools might differ for various stakeholders. For instance, an AI expert trained in this field can understand and interpret the reasoning behind an AI algorithm better and quicker than a nonexpert in AI. Therefore, health care providers and patient education about AI can lead to a better understanding of the algorithm, which leads to a better understanding of the explainability of an AI intervention. In brief, end users' understanding of the predictions/decisions made by the AI intervention, as well as increased explainability and interpretability of the AI tool, can increase end-user *trust* in the outcome given [57].

A lack of trustworthiness is one of the many bioethical barriers that may arise when implementing an AI intervention in health care and SDM; therefore, improving AI literacy in both patients and health care providers, as well as increasing the explainability and interpretability of AI systems, trust can be increased. In addition, there is a discrepancy in the literature regarding the level of explainability required within the health care setting to ensure a proper understanding of and trust in the outcomes provided by the algorithm [58]. Future studies are required to determine how to efficiently educate end users about AI-SDM tools, how to efficiently incorporate explainability and interpretability in this context, and how much explainability and interpretability are deemed sufficient in this context and the context of informed consent.

Human-Centered AI

Of the included articles, 3 [39,41,44] involved health care providers in the validation stage of the AI system, and 1 included both health care providers and patients in the clinical validation stage of the AI system [43]. One article [38] included patients and health care providers in co-design activities, resulting in user-generated versions of the developed tool. However, no details were provided on how the co-design activity was organized, and end users were not involved in the subsequent design and development of the AI tool.

Further efforts are needed, both from the AI and SDM communities, to include health care providers and patients (as end users of the developed AI systems) in the design, development, validation, and implementation of AI-SDM tools. SDM is the core of patient-centered care; thus, patient values and preferences need to be considered in every step defining the process. Ethicists argue that by not using patient preferences or values as input or influencing the output, but rather leaving the *shared decision* aspect to the patient choosing from evidence-based options presented by the AI, the process is not truly patient centered [59].

Thus, to ensure that SDM fundamentally occurs when AI interventions are introduced, patient preferences must be incorporated into the design. Termed *value-sensitive design*, this method incorporates human values throughout the design process [59]. However, the successful incorporation of individual patient values into algorithm design and how to efficiently include patients and health care providers in the development and validation of AI systems in health for SDM remains a challenge, and further studies are required. A recent assessment of the current methods showed that most existing user-centered design methods were primarily created for non-AI systems and did not effectively address the unique issues in AI systems [60]. This is also the case for AI-SDM tools.

Reporting on AI Interventions

In our review, we observed poor reporting of AI interventions in the included studies. Studies that report AI interventions should use validated frameworks and guidelines to report their results. Transparent and complete reporting of AI interventions supporting SDM is important for detecting errors and potential biases and evaluating the usefulness of the intervention [61]. An example of such a reporting framework is the Transparent Reporting of a multivariable prediction model of Individual Prognosis or Diagnosis (TRIPOD), which consists of a checklist of items deemed essential for transparent reporting [62]. As the original framework is primarily applied to regression-based predictive models, the TRIPOD-AI extension is being developed, specifically for machine learning–based prediction model studies [63]. Transparent and complete reporting allows for a good understanding and encourages reproducibility of the work in future studies, which is an important factor to consider in the growing implementation of AI-SDM in clinical settings.

None of the articles included in this review mentioned adhering to a specific reporting framework or considered reproducibility. This resulted in a lack of clarity in the included articles regarding different aspects, including whether the training data set was representative, how the potential bias (eg, representativeness and algorithmic biases) and missing data were considered, how AI had been used in the clinical setting, and what were the outcomes resulting from AI implementation. In fact, only 1 article [44] comprehensively reported on the sociodemographic characteristics of the participants involved in the use of AI intervention. Such reporting should be standardized so that AI interventions and clinical implementations can be better understood and compared effectively. The importance of using a reporting framework needs to be emphasized in future AI studies to promote an increased understanding and reproducibility of AI-SDM in clinical contexts.

Limitations of the Study

We did not conduct a quality appraisal of the included articles, although it is not common, nor is it required to include within a scoping review. However, our review sheds light on this important area, and there are some areas for improvement. Our inclusion criteria were quite strict, and only included articles in which AI intervention was used to support the decision-making point in SDM. Therefore, we may have missed work related to other aspects of SDM. Further systematic reviews may be needed in this area to ensure that the results of this review can be applied in policy and practice.

Conclusions

In this scoping review, we demonstrated the extent and variety of AI systems being tested and implemented in SDM, showed that this field is expanding, and highlighted that knowledge gaps remain and should be prioritized in future studies. Our findings suggest that existing evidence on the use of AI to support SDM is in its infancy. The low number of included studies shows that not much research has been conducted to test, implement, and evaluate the impact of AI on SDM. Future research is required to strengthen and standardize the use of AI intervention in different steps of SDM and to evaluate its impact on particular decisions, populations, and settings. Greater focus and effort from the research community needs to be made on addressing the aspects of explainability, interpretability, reproducibility, and human-centered AI, especially when developing an intervention of their own. Finally, future research should further investigate which SDM steps will benefit most from what type of AI and how AI interventions can be applied to enforce the patient–health care provider relationship.

Acknowledgments

This study was funded by a start-up fund from McGill University (principal investigator: SAR). The authors would like to acknowledge this support. SAR receives salary support, that is, Research Scholar Junior 1 Career Development Award, from the Fonds de Recherche du Québec-Santé, and her research program is supported by the Natural Sciences and Engineering Research Council (Discovery Grant 2020-05246). FL is tier 1 Canada Research Chair in Shared Decision-Making and Knowledge Translation. The authors thank Milad Ghanbari, Sara Makaremi, and Stewart McLennan for their contribution to this work. The authors also thank the Quebec SPOR SUPPORT (Support for People and Patient-Oriented Research and Trials) Unit for their methodological support.

Authors' Contributions

The authors have reported the contributions according to the Contributor Roles Taxonomy. SAR and PP contributed to conceptualization. SAR, RG, PP, HTVZ, and GG contributed to the methodology. SAR and MC contributed to data curation. SAR, YH, PG, and MC contributed to the formal analysis (see the Acknowledgments section). SAR contributed to funding acquisition, project administration, and resources. SAR, YH, and GG contributed to the investigation. SAR and MC wrote the original draft of this paper. SAR, YH, PG, MC, RG, GG, HTVZ, FL, PP, and DP contributed to reviewing and editing the article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist containing the page number where each reporting criterion is addressed.
[DOCX File , 38 KB - medinform_v10i8e36199_app1.docx]

Multimedia Appendix 2
Detailed data extraction table.
[DOCX File , 22 KB - medinform_v10i8e36199_app2.docx]

References

1. Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). Soc Sci Med 1997 Mar;44(5):681-692. [doi: 10.1016/s0277-9536(96)00221-3]
2. Barry MJ, Edgman-Levitan S. Shared decision making — the Pinnacle of patient-centered care. N Engl J Med 2012 Mar;366(9):780-781. [doi: 10.1056/nejmp1109283]
3. Couët N, Desroches S, Robitaille H, Vaillancourt H, Leblanc A, Turcotte S, et al. Assessments of the extent to which health-care providers involve patients in decision making: a systematic review of studies using the OPTION instrument. Health Expect 2015 Aug 04;18(4):542-561 [FREE Full text] [doi: 10.1111/hex.12054] [Medline: 23451939]
4. Edwards M, Davies M, Edwards A. What are the external influences on information exchange and shared decision-making in healthcare consultations: a meta-synthesis of the literature. Patient Educ Couns 2009 Apr;75(1):37-52. [doi: 10.1016/j.pec.2008.09.025] [Medline: 19036550]

5. Holmes-Rovner M, Valade D, Orlowski C, Draus C, Nabozny-Valerio B, Keiser S. Implementing shared decision-making in routine practice: barriers and opportunities. Health Expect 2000 Sep;3(3):182-191 [FREE Full text] [doi: 10.1046/j.1369-6513.2000.00093.x] [Medline: 11281928]
6. Légaré F, Ratté S, Gravel K, Graham ID. Barriers and facilitators to implementing shared decision-making in clinical practice: update of a systematic review of health professionals' perceptions. Patient Educ Couns 2008 Dec;73(3):526-535. [doi: 10.1016/j.pec.2008.07.018] [Medline: 18752915]
7. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct 23;27(10):1361-1367 [FREE Full text] [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
8. Elwyn G, Durand MA, Song J, Aarts J, Barr PJ, Berger Z, et al. A three-talk model for shared decision making: multistage consultation process. BMJ 2017 Nov 06;359:j4891 [FREE Full text] [doi: 10.1136/bmj.j4891] [Medline: 29109079]
9. McCarthy J. What is Artificial Intelligence?. Cambridge, Massachusetts, United States: MIT press; 1997.
10. Amisha, Malik P, Pathania M, Rathaur V. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019 Jul;8(7):2328-2331 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
11. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect 2020 May;26(5):584-595 [FREE Full text] [doi: 10.1016/j.cmi.2019.09.009] [Medline: 31539636]
12. Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev 1959 Jul;3(3):210-229. [doi: 10.1147/rd.33.0210]
13. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HT, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res 2021 Sep 03;23(9):e29839 [FREE Full text] [doi: 10.2196/29839] [Medline: 34477556]
14. Bi Q, Goodman K, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol 2019 Dec 31;188(12):2222-2239. [doi: 10.1093/aje/kwz189] [Medline: 31509183]
15. Balmer R, Levin S, Schmidt S. Artificial intelligence applications in telecommunications and other network industries. Telecommun Policy 2020 Jul 23;44(6):101977 [FREE Full text] [doi: 10.1016/j.telpol.2020.101977]
16. Arslanian H, Fischer F. The Future of Finance The Impact of FinTech, AI, and Crypto on Financial Services. Cham: Springer International Publishing; 2019.
17. Topol E. Deep Medicine How Artificial Intelligence Can Make Healthcare Human Again. New York, United States: Basic Books; 2019.
18. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug 17;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
19. Ting DS, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb 25;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
20. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018 Jun 12;71(23):2668-2679 [FREE Full text] [doi: 10.1016/j.jacc.2018.03.521] [Medline: 29880128]
21. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017 Dec 06;88(6):581-586 [FREE Full text] [doi: 10.1080/17453674.2017.1344459] [Medline: 28681679]
22. Niazi MK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019 May;20(5):e253-e261. [doi: 10.1016/s1470-2045(19)30154-8]
23. Wang X, Peng X, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017; Honolulu, HI, USA. [doi: 10.1109/cvpr.2017.369]
24. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219. [doi: 10.1056/nejmp1606181]
25. Kohane IS, Drazen JM, Campion EW. A glimpse of the next 100 years in medicine. N Engl J Med 2012 Dec 27;367(26):2538-2539. [doi: 10.1056/nejme1213371]
26. 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]
27. Triberti S, Durosini I, Pravettoni G. A "Third Wheel" effect in health decision making involving artificial entities: a psychological perspective. Front Public Health 2020 Apr 28;8:117 [FREE Full text] [doi: 10.3389/fpubh.2020.00117] [Medline: 32411641]
28. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics 2020 Apr 03;47(12):e3 [FREE Full text] [doi: 10.1136/medethics-2019-105860] [Medline: 32245804]

29. Hassan N, Slight RD, Bimpong K, Weiand D, Vellinga A, Morgan G, et al. Clinicians' and patients' perceptions of the use of artificial intelligence decision aids to inform shared decision making: a systematic review. Lancet 2021 Nov;398:S80. [doi: 10.1016/s0140-6736(21)02623-4]
30. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
31. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
32. The Joanna Briggs Institute Reviewers' Manual 2015 Methodology for JBI Scoping Reviews. South Australia: The Joanna Briggs Institute; 2015.
33. Artificial intelligence supporting shared decision making: a scoping review. OSF. URL: https://osf.io/dwzbf/ [accessed 2021-12-04]
34. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
35. Stone PW. Popping the (PICO) question in research and evidence-based practice. Appl Nurs Res 2002 Aug;15(3):197-198. [doi: 10.1053/apnr.2002.34181] [Medline: 12173172]
36. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. London, United Kingdom: Pearson; 2002.
37. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product From the ESRC Methods Programme. United Kingdom: Lancaster University; 2006. [doi: 10.13140/2.1.1018.4643]
38. Kökciyan N, Chapman M, Balatsoukas P, Sassoon I, Essers K, Ashworth M, et al. A collaborative decision support tool for managing chronic conditions. Stud Health Technol Inform 2019 Aug 21;264:644-648. [doi: 10.3233/SHTI190302] [Medline: 31438003]
39. Kokciyan N, Sassoon I, Sklar E, Modgil S, Parsons S. Applying metalevel argumentation frameworks to support medical decision making. IEEE Intell Syst 2021 Mar 1;36(2):64-71. [doi: 10.1109/mis.2021.3051420]
40. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
41. Frize M, Yang L, Walker R, O'Connor A. Conceptual framework of knowledge management for ethical decision-making support in neonatal intensive care. IEEE Trans Inf Technol Biomed 2005 Jun;9(2):205-215. [doi: 10.1109/titb.2005.847187] [Medline: 16138537]
42. Wang Y, Li P, Tian Y, Ren J, Li J. A shared decision-making system for diabetes medication choice utilizing electronic health record data. IEEE J Biomed Health Inform 2017 Sep;21(5):1280-1287. [doi: 10.1109/jbhi.2016.2614991]
43. Twiggs JG, Wakelin EA, Fritsch BA, Liu DW, Solomon MI, Parker DA, et al. Clinical and statistical validation of a probabilistic prediction tool of total knee arthroplasty outcome. J Arthroplasty 2019 Nov;34(11):2624-2631. [doi: 10.1016/j.arth.2019.06.007] [Medline: 31262622]
44. Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. JAMA Netw Open 2021 Feb 01;4(2):e2037107 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.37107] [Medline: 33599773]
45. Siau K, Wang W. Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technol J 2018;31(2):47.
46. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Artificial Intelligence in Healthcare. Cambridge, Massachusetts, United States: Academic Press; 2020. [doi: 10.1016/B978-0-12-818438-7.00012-5]
47. Debnath S, Barnaby DP, Coppa K, Makhnevich A, Kim EJ, Chatterjee S, Northwell COVID-19 Research Consortium. Machine learning to assist clinical decision-making during the COVID-19 pandemic. Bioelectron Med 2020 Jul 10;6(1):14 [FREE Full text] [doi: 10.1186/s42234-020-00050-8] [Medline: 32665967]
48. Association for the Advancement of Artificial Intelligence. AI and health communication. In: Proceedings of the 2011 AAAI Spring Symposium, Technical Report SS-11-01. 2011 Presented at: 2011 AAAI Spring Symposium; Mar 21-23, 2011; Stanford, California, USA.
49. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021 Sep 21;4(1):140 [FREE Full text] [doi: 10.1038/s41746-021-00509-1] [Medline: 34548621]
50. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020 Jun 23;58:82-115 [FREE Full text] [doi: 10.1016/j.inffus.2019.12.012]
51. Holzinger A, Malle B, Saranti A, Pfeifer B. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. Inf Fusion 2021 Jul 23;71:28-37. [doi: 10.1016/j.inffus.2021.01.008]
52. The European legal framework for medical AI. In: Machine Learning and Knowledge Extraction. Cham: Springer; 2020.

53. Shulman S. Survey shows next era of healthcare will be powered by AI. Intel. URL: https://www.intel.com/content/www/us/en/newsroom/opinion/survey-shows-next-era-healthcare-powered-ai.html [accessed 2021-12-05]
54. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 04;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
55. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. JAMA 2018 Sep 18;320(11):1107-1108. [doi: 10.1001/jama.2018.11029] [Medline: 30178025]
56. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. JAMA 2019 Aug 13;322(6):497-498. [doi: 10.1001/jama.2018.20563] [Medline: 31305873]
57. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 2020 Jul 31;11(1):3852 [FREE Full text] [doi: 10.1038/s41467-020-17431-x] [Medline: 32737308]
58. Diprose W, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J Am Med Inform Assoc 2020 Apr 01;27(4):592-600 [FREE Full text] [doi: 10.1093/jamia/ocz229] [Medline: 32106285]
59. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics 2019 Mar 22;45(3):156-160. [doi: 10.1136/medethics-2018-105118] [Medline: 30467198]
60. Xu W, Dainoff MJ, Ge L, Gao Z. Transitioning to human interaction with AI systems: new challenges and opportunities for HCI professionals to enable human-centered AI. Int J Human Comput Interact 2022 Apr 06:1-25. [doi: 10.1080/10447318.2022.2041900]
61. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. Lancet 2019 Apr;393(10181):1577-1579. [doi: 10.1016/s0140-6736(19)30037-6]
62. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 07;350(jan07 4):g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]
63. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 09;11(7):e048008 [FREE Full text] [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analysis extension for Scoping Reviews
SDM: shared decision making
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 05.01.22; peer-reviewed by I Ramos-Herrera, T Agoritsas, T Gladman; comments to author 26.02.22; revised version received 16.04.22; accepted 21.04.22; published 09.08.22.

Please cite as:
Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, Ghadiri P, Grad R, Poenaru D, Gore G, Zomahoun HTV, Légaré F, Pluye P
Application of Artificial Intelligence in Shared Decision Making: Scoping Review
JMIR Med Inform 2022;10(8):e36199
URL: https://medinform.jmir.org/2022/8/e36199
doi:10.2196/36199
PMID:35943793

Viewpoint

Twenty Years of the Health Insurance Portability and Accountability Act Safe Harbor Provision: Unsolved Challenges and Ways Forward

Brittany Krzyzanowski[1], PhD; Steven M Manson[1], PhD

University of Minnesota, Minneapolis, MN, United States

Corresponding Author:
Brittany Krzyzanowski, PhD
University of Minnesota
269 19th Avenue South
Minneapolis, MN, 55455
United States
Phone: 1 612 625 5000
Email: krzyz016@umn.edu

Abstract

The Health Insurance Portability and Accountability Act (HIPAA) was an important milestone in protecting the privacy of patient data; however, the HIPAA provisions specific to geographic data remain vague and hinder the ways in which epidemiologists and geographers use and share spatial health data. The literature on spatial health and select legal and official guidance documents present scholars with ambiguous guidelines that have led to the use and propagation of multiple interpretations of a single HIPAA safe harbor provision specific to geographic data. Misinterpretation of this standard has resulted in many entities sharing data at overly conservative levels, whereas others offer definitions of safe harbors that potentially put patient data at risk. To promote understanding of, and adherence to, the safe harbor rule, this paper reviews the HIPAA law from its creation to the present day, elucidating common misconceptions and presenting straightforward guidance to scholars. We focus on the 20,000-person population threshold and the 3-digit zip code stipulation of safe harbors, which are central to the confusion surrounding how patient location data can be shared. A comprehensive examination of these 2 stipulations, which integrates various expert perspectives and relevant studies, reveals how alternative methods for safe harbors can offer researchers better data and better data protection. Much has changed in the 20 years since the introduction of the safe harbor provision; however, it continues to be the primary source of guidance (and frustration) for researchers trying to share maps, leaving many waiting for these rules to be revised in accordance with the times.

(JMIR Med Inform 2022;10(8):e37756) doi:10.2196/37756

KEYWORDS

Health Insurance Portability and Accountability Act; HIPAA; data privacy; health; maps; safe harbor; visualization; patient privacy

Introduction

Background

When addressing many types of research problems, maps should generally be shared at a resolution that best portrays the reality of the underlying data. In terms of health and disease mapping, this realism often means desiring a fine-detailed visualization that helps make community-level public health interventions more effective. Geotechnology offers innovative ways of creating these fine-detailed maps and customizing them for the analysis and display of health data. However, at the same time, these data and tools can be dangerous when working with sensitive data, such as patient health records. In particular, scholars must be careful not to share maps that contain so much detail that individuals can be identified. To prevent the identification of patient records, in the United States, the Health Insurance Portability and Accountability Act (HIPAA) provides guidance on ways of deidentifying protected health information (PHI) before it is shared; however, HIPAA guidelines are difficult to apply to spatial data.

The HIPAA law poses several challenges to researchers seeking to use and share spatial data. First, many researchers find core elements of the *safe harbor provisions* of HIPAA (a set of conditions that define how data can be shared) ambiguous or difficult to understand, which is reflected in the disagreement and uncertainty in research and policy circles on how to meet

the safe harbor standards. Second, playing it safe by taking a conservative approach to sharing maps to better meet the safe harbor standard—most often by releasing only highly aggregated maps or no maps at all—is a form of data loss that imposes potentially serious costs as it does not allow for the examination of local health distributions at reasonable resolutions for many common health problems. These 2 challenges lead to disagreement on how to follow privacy rules, and in fact, many scholars and policy makers have challenged these rules, saying that it is possible to share finer-grained mapped health data without jeopardizing patient privacy.

Addressing the twin challenges of the safe harbor provisions (ambiguity and data loss) requires an exploration of past and current understanding of how the provisions are enacted and identification of specific ways in which finer-scaled data may be legally and technically possible. The following section of this paper begins this exploration by examining the legal dimensions of the HIPAA law, from its creation to current practice. This section examines the events and concerns that fueled the motivations of those who helped write the safe harbor provisions, with a particular focus on answering the question of why zip codes and a population threshold of 20,000 were chosen as anchors for the safe harbors. The following section explores the first of the twin challenges—uncertainty—and establishes how some unintentional ambiguity in the law has led to different interpretations of HIPAA privacy provisions specific to geographic data in the public health literature. We focus on how this ambiguity has led to 2 common but different interpretations across a range of scholarships based on 3-digit and 5-digit zip codes and what this means for mapped data. The following section presents and explores data loss, the second of the twin challenges of the safe harbor provisions. The section builds on the previous ones to explore whether there is a middle ground between sufficiency and stringency, asking, in essence, if there are ways of minimizing risk under HIPAA while allowing for more useful maps. This paper concludes by presenting new approaches to the deidentification of patient data and discusses ways forward.

This paper advances our understanding, and potential use, of the safe harbor provision of HIPAA law, as applied to spatial data presented as maps. It is the first comprehensive overview of the long-standing and important conversations on this general topic. By untangling the law and reviewing its history and use, this paper offers avenues for finding safe and more useful ways of sharing mapped patient data. In addition, it seeks to spur a broader conversation on ways forward that necessarily expand and improve shared understanding of privacy regulations to encourage researchers to investigate alternative strategies.

HIPAA Privacy Act: Zip Codes and the 20,000-Person Population Threshold

Overview

To better understand the safe harbor provision and what it asks of researchers, it is best to first understand its origin. Examining HIPAA in terms of its history and evolution sheds light on how to approach the sharing of geographic information under the safe harbor standard. We asked two related questions: (1) why do zip codes hold such sway over defining the safe harbor rule, and (2) why is a threshold of 20,000 people used to define privacy? Answering these questions clarifies some of the key ambiguities in HIPAA safe harbors and provides insight into why there is so much seeming disagreement within and across research domains. The following section provides a brief overview of HIPAA privacy law before diving into the history of the safe harbor provision to provide insights into the 2 key ambiguities (the use of zip codes and the population threshold).

The Safe Harbor Provision

To protect patient privacy, HIPAA limits the ways in which patient data can be shared. Patient data are considered PHI that needs to be kept secure as they include private medical information along with identifying information such as names, birth dates, addresses, and social security numbers. Address data, in particular, are considered extremely sensitive as they (along with other location data such as longitude and latitude) may be used to pinpoint the residence of an individual. This degree of locational specificity substantially increases the likelihood of identification, if not fully guaranteeing identification in the case of single-occupant residences. For this reason, patient locations need to be masked in accordance with HIPAA privacy law.

Two standards are specified under the HIPAA rule for deidentifying patient data—the safe harbor standard and expert determination—but the former is the de facto standard [1]. Expert determination—also termed as the statistical standard—is the process by which an investigator masks their data and has a third-party expert determine whether the applied location masking strategy provides a low probability of identification [1]. Expert determination is not frequently used in large part as it is ambiguous and requires unspecified documentation, in addition to placing a great deal of pressure on the third-party expert who is charged with certifying HIPAA compliance. This leaves the safe harbor standard as the most commonly relied upon practice for deidentifying patient data [2]. Its immediate appeal, and the primary reason for broader acceptance than expert determination, is that it offers ostensibly clear guidance. The safe harbor standard is the focus of the remainder of this paper.

In essence, the safe harbor method protects patient data by simply removing 18 types of identifiers (Textbox 1). Many of these elements are straightforward to comprehend and implement, such as not including names, birth dates, and social security numbers. Some of the other elements pose their own challenges in an age of surveillance, such as biometric markers, including vehicle license plates and facial imagery. However, our focus is section 2 of the safe harbor relating to the patient's location, which is especially relevant to mapping and, not surprisingly, the primary source of confusion in applying the safe harbor rule to mapping. The location provision of the safe harbor rule requires a minimum population of at least 20,000 people to be contained within each aggregated geographical unit, and the rule further requires that the only permissible geography (smaller than the state) be a form of zip code.

Ambiguity arises when the type of zip code is not specified. Although it seems fairly clear from Textbox 1 that the rule intends for investigators to rely on the use of 3-digit zip codes (compared with 5-digit zip codes), not all who read this stipulation see it that way. There are many reasons for this, including various misleading representations of the rule found in legal web-based documentation and in the literature on public health and disease mapping [3-11]. The following section explores how zip codes have come to play a key role in the safe harbor rule.

Textbox 1. The key elements of the safe harbor provision.

The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

1. Names
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial 3 digits of the zip code if, according to the current publicly available data from the Bureau of the Census, the geographic unit formed by combining all zip codes with the same 3 initial digits contains >20,000 people, and the initial 3 digits of a zip code for all such geographic units containing ≤20,000 is changed to 000
3. All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages >89 years, and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of the age of ≥90 years
4. Telephone numbers
5. Vehicle identifiers and serial numbers, including license plate numbers
6. Fax numbers
7. Device identifiers and serial numbers
8. Email addresses
9. Web Universal Resource Locators (URLs)
10. Social security numbers
11. IP addresses
12. Medical record numbers
13. Biometric identifiers, including finger and voice prints
14. Health plan beneficiary numbers
15. Full-face photographs and any comparable images
16. Account numbers
17. Any other unique identifying number, characteristic, or code, except as permitted by paragraph c of this section (paragraph c is presented in the section “Re-identification”)
18. Certificate and license numbers

Why Zip Codes?

If we were to remove zip codes from the safe harbor provision, there would be no ambiguity in terms of its interpretation as the rule would simply focus on the threshold of 20,000 people to define whether an arbitrary geographical unit is sufficient. Hence, why are zip codes still written into the law? To answer this, we need to start at the very beginning and understand how the political, social, and technological milieu of the early and mid-1990s shaped some core principles and guidelines. Zip codes were originally not included in the rule; however, this quickly changed as a result of a mix of happenstance and deliberation. The following paragraphs provide insight into the series of events that led to the HIPAA safe harbor provision that we understand today, beginning with the proposed bill.

Before HIPAA was law, it was a bill, specifically bill *H.R. 3103* of the 104th Congress from 1995 to 1996. This bill was introduced in the spring of 1996 as part of an initial attempt at health care reform by the Clinton administration. The overarching focus of H.R. 3103 was to improve access to health care and address fraud, waste, and abuse in health insurance and health care delivery; however, it also—quite briefly—mentions a specific interest in the protection of patient data (section 1177 of H.R. 3103, 1996). In a single paragraph, the bill addresses the wrongful disclosure of individually identifiable health information, in large part, as it relates to insurance fraud and abuse:

> *A person who knowingly and in violation of this part uses or causes to be used a unique health identifier; obtains individually identifiable health information relating to an individual; or discloses individually identifiable health information to another person, shall...be fined not more than $50,000, imprisoned not more than 1 year, or both; if the offense is committed under false pretenses, be fined not more than $100,000, imprisoned not more than 5 years, or both; and if the offense is committed with intent to sell, transfer, or use individually identifiable health*

> *information for commercial advantage, personal gain, or malicious harm, fined not more than $250,000, imprisoned not more than 10 years, or both.* [Section 1177. Wrongful disclosure of individually identifiable health information]

This bill was the first step toward the development of a series of protections that would eventually become the HIPAA privacy law that we know today. However, much changed during the journey from the bill's initial proposal to the passage of the final law and attendant guidelines, especially in terms of modifications made to the data privacy and deidentification standards. Early renditions of HIPAA provided very little guidance on how to define deidentified health information. Mass computerization of individual health information had only just begun, with electronic health records making their first appearance in 1992 [12]. In the mid-1990s, with the rise of the internet and home computers, threats to data privacy elicited much fear among the American public [13]. Despite these concerns, when the bill went to Congress in the summer of 1996, the disclosure of identifiable health information was not documented as a part of the discussion on the congressional record [14].

A year after its introduction, Sweeney [15], a computer scientist working at the Massachusetts Institute of Technology, purchased a voter registration list for Cambridge, Massachusetts, United States, and cross-referenced it with a "de-identified" (meaning the names were missing but other information such as birth date remained) Massachusetts Group Insurance hospitalization data set that was provided to researchers. Sweeney [15] determined that by using birth date, gender, and a 5-digit zip code, she could match a patient's medical records with their name on the voter registration list. This meant that for only US $20 (the cost of the voter registration list), Sweeney [15] could *potentially* identify (by name) some of the registered voters and their medical records, which included sensitive information such as diagnoses, procedures, and medications. With this knowledge, Sweeney [15] famously mailed the governor of Massachusetts his own medical records. This event fueled anxiety about the potential misuse of patient information and put data protection at the forefront of many conversations on privacy reform. The study by Sweeney [15] was central to the next chapter of the story of HIPAA's evolution, the 1999 Notice of Proposed Rulemaking (NPRM) [16,17].

In response to the work by Sweeney [15], the 1999 NPRM proposed a stringent definition of deidentified health information. Of particular interest to this paper is how the NPRM defined the smallest unit of allowable geography as the state. All other geographic identifiers would be removed, meaning that street addresses, cities, counties, and both 3- and 5-digit zips were not permissible. This state-level geographic standard was too restrictive for any researcher interested in studying the geographic variation in health and disease, such as geographers and epidemiologists. Under such rules, researchers are only able to publish maps at the state level (usually at the national level). For most scholars, this limit meant that only statistical point estimates (such as regression output) could be published under the safe harbor rule.

Fortunately, for researchers, feedback from the 1999 NPRM's call for public comments pushed the Department of Health and Human Services (HHS) to allow slightly more geographic information to be shared as deidentified information. The safe harbor standard's 3-digit zip code rule made its first appearance on a federal record [18]. The rule states the following:

> *In the safe harbor, we explicitly allow...some geographic location information to be included in the deidentified information, but...zip codes must be removed or aggregated (in the form of most three-digit zip codes) to include at least 20,000 people.*

Compared with the 1999 NPRM guidelines, this safe harbor standard was much less stringent but still meant to withstand a population-level identification attack of the sort developed by Sweeney [15], which required 5-digit zip codes.

This simple 3-digit zip code rule became more complicated in the decade after HIPAA was promulgated. The initial formulation seemed clear (3-digit zip codes were the intended level of aggregation); however, subsequent modifications to HIPAA introduced ambiguity. Changes to the final rule in 2002 left out the key clause that made it clear that 3-digit zip codes would be the *only* permissible form of aggregation (other than the state level) [19]. This contributed to the ever-growing ambiguity regarding the provision of geographic deidentification, and along with other nebulous aspects of the law, many researchers found it difficult to navigate HIPAA. Therefore, with the passage of the Health Information Technology for Economic and Clinical Health Act in 2009, the HHS was required "to issue guidance on methods for de-identification of PHI as designated in HIPAA's Privacy Rule." In response, the US Office of Civil Rights (OCR) held a workshop in 2010 to provide guidance on strategies for the deidentification of PHI. OCR used input from panelists, including Sweeney and Barth-Jones (noted later in this paper), and workshop attendees to develop a lengthy guidance document [1]. This comprehensive document is helpful in that it provides a more detailed description of the safe harbor rule; however, unfortunately, it still contained the same ambiguous phrasing (regarding zip codes) found in the modifications of the written law. To make matters worse, the landing page for the workshop on HIPAA's deidentification standard (which features a link to the guidance document page) uses the term *geocodes* rather than zip codes (Textbox 2 provides the full phrasing) when referring to aggregating geographic data, which could easily lead readers to believe that any unit (not only zip codes) could be used for aggregation. These ambiguities, alongside inconsistencies in use and opinion found throughout the literature (explored below in section *Twin Challenge 1: Ambiguity*) about core HIPAA documents [1,19], may have contributed to the widespread confusion that continues today.

Textbox 2. The various ways investigators interpret the geographic location stipulation of the Health Insurance Portability and Accountability Act (HIPAA) safe harbor rule.

Paper, author, and interpretation

- Confidentiality risks in fine scale aggregations of health data (Curtis et al [6])
  - "Unfortunately there are few guidelines with regards the release of aggregated data. A commonly discussed threshold between researchers is that health data should only be visualized for ZIP codes with a base population of no less than 20,000."
- Reidentification risks in HIPAA safe harbor data: a study of data from one environmental health study (Sweeney et al [10,20])
  - "[T]he provision requires removing explicit identifiers (such as name, address and other personally identifiable information), reporting dates in years, and reducing some or all digits of a postal (or ZIP) code."
- Workshop on the HIPAA privacy rule's deidentification standard (US Office of Civil Rights [11])
  - "[The Safe Harbor approach] permits a covered entity to consider data to be de-identified if it removes 18 types of identifiers (eg, names, dates, and geocodes on populations with less than 20,000 inhabitants) and has no actual knowledge that the remaining information could be used to identify an individual, either alone or in combination with other information."
- Conforming to HIPAA regulations and compilation of research data (Clause et al [3])
  - "Implementation of these methods can be somewhat difficult for the clinical researcher for data sets of less than 20,000 records (as determined by collapsing populated geographic codes representing sparse populations)."
- From healthy start to hurricane Katrina: using GIS to eliminate disparities in perinatal health (Curtis [4])
  - "The error of recording '70808' rather than '70806' in Baton Rouge would involve considerable changes in social, economic, and racial contexts. This is a problem if data are only available by zip code, which unfortunately is still too common in terms of releasing data for GIS analysis."
  - "Although there are HIPAA regulations regarding the display of aggregate data on choropleth maps, these guidelines are generally considered too restrictive for useful cartography (only zip codes with more than 20 000 can be visualized)."
- A linear programming model for preserving privacy when disclosing patient spatial information for secondary purposes (Jung and El Emam [7])
  - "A prevailing method to create de-identified data sets is to aggregate pre-defined areas, such as ZIP codes or counties, into a new area."
  - "Yet, the first three digits of a ZIP code may be included, provided that at least 20,000 people share the same first three digits."
- The challenges of creating a gold standard for deidentification research (Browne et al [8])
  - "[The guidelines of the Privacy Rule] say that units smaller than a state should be redacted, although Baltimore has a population of well over 20,000, the size limit for Zip-Codes. D.C. was considered a state for this purpose."
- Challenges and insights in using HIPAA privacy rule for clinical text annotation (Kayaalp et al [9])
  - "The Privacy Rule states that information about all geographic subdivisions smaller than state, except the first two digits of the zip code, must be de-identified. The third digit of the zip code can be left intact, only if the size of the population in the area of the censored two digits is greater than 20,000 according to the most recent census data."
- Broken promise of privacy: responding to the surprising failure of anonymization (Ohm [5])
  - "Id. § 164.514(b)(2)(B) (allowing only two digits for ZIP codes with 20,000 or fewer residents)."

Why 20,000 People?

Part of the ambiguity surrounding the use of zip codes is tied to the 20,000-person threshold in defining safe harbor rules. The decision to allow substate-level geographies, specifically zip codes, is partially tied to research on the role of the population size in protecting privacy. In simple terms, by increasing the number of people reported within a given region, the chances of successfully matching an individual in that region to their health records decreases. This is because the odds of a unique combination of identifying characteristics occurring in a population decline as the number of people in a data set increases.

How did the HHS determine that 20,000 was the appropriate population threshold? To answer this, we must look to the proposed final rule [18] as there is little to no discussion of this determination within the literature or on the HHS support and guidance webpages. In the final rule, the HHS points to the precedent of how the Bureau of the Census "shares geographical units only if they contain populations of at least 100,000 people" [20]. This standard is conservative, and thus, the HHS turned to other sources so that they might be able to drop the threshold lower.

Specifically, the HHS drew on 2 simulation studies, one by Greenberg and Voshell [21] and the second by Horm [22]. These studies explored how the proportion of unique records within

a data set can be influenced by changes in the size of the population and the number and type of variables included. For instance, approximately 7.3% of records within the 1990 census are unique, or potentially identifiable, given the 100,000-person population threshold using standard census variables such as age, race, ethnicity, sex, and housing or household information [23]. Nevertheless, the proportion of unique records is a function of the available information. Sharing a greater number of variables increases the potential to identify an individual; therefore, the Census Bureau population threshold increases from 100,000 to ≥250,000 when greater numbers of variables are released as microdata [20].

There is a point at which increasing the size of the population no longer adds notable increases to data protection. For census data, when only 6 demographic variables are shared, there is a point of diminishing returns for approximately 20,000 people [21]. In addition to the number of demographic variables, the type of variables shared also matters. For instance, a population of 25,000 contains 25% unique records when 9 variables are shared; however, when the occupation variable is removed, this proportion drops to 10% [22]. In this case, occupation can be particularly identified, given that some occupations are much rarer than others. The HHS drew on this scholarship to make their determination [23]:

> *After evaluating current practices and recognizing the expressed need for some geographic indicators in otherwise de-identified databases, we concluded that permitting geographic identifiers that define populations of greater than 20,000 individuals is an appropriate standard that balances privacy interests against desirable uses of de-identified data. In making this determination, we focused on the studies by the Bureau of Census cited above which seemed to indicate that a population size of 20,000 was an appropriate cut off if there were relatively few (6) demographic variables in the database. Our belief is that, after removing the required identifiers to meet the safe harbor standards, the number of demographic variables retained in the databases will be relatively small, so that it is appropriate to accept a relatively low number as a minimum geographic size.*

In addition, as the HHS considers the 20,000-person population stipulation, the lowest bound could also be tied to the adoption of the 3-digit zip. Although 3-digit zip codes vary widely in terms of the size of the population they contain (in 2020, ranging from 3147 to 3,310,455 people), only 18 zip codes of 3 digits containing <20,000 people at the time the safe harbor was first determined. Currently, there are only 13 zip codes of 3 digits in the nation, which are too small and would need to be merged with neighboring geographies to meet the minimum threshold of 20,000 people [24]. Fortunately, as most 3-digit zip codes contain populations of >20,000 people, researchers following the 3-digit zip code rule are not often burdened with the task of data aggregation. Perhaps the HHS hoped that using these 3-digit zip codes could help enforce a more conservative following of the population threshold while also making the guidelines more straightforward. Unfortunately, this is not the case in many important ways.

Twin Challenge 1: Ambiguity

Overview

The safe harbor rule seems straightforward when seen from the original final rule of 2000; however, given the modifications, as well as how it appears in the literature today, it carries an essential ambiguity that has led to large gaps and disagreements in research and policy work. We first examine different interpretations of the rule based on these ambiguities and draw examples from the scientific literature to show how different scholars rely on different interpretations. We then simplify the discussion by proposing that the crux of many disagreements—and the basis of productive ways forward—can be seen by focusing on the use of 3-digit and 5-digit zip codes.

Safe Harbor Provision and Zip Code Ambiguity

The primary driver of disagreements in the literature seems to hinge on how individual researchers and teams interpret the role of zip codes versus the 20,000-person threshold. This often comes to the fore in determining how much location data must be removed from patient data to satisfy HIPAA requirements.

The potential for misunderstanding stems from one part of the provision—the piece regarding geographic information that states the following with respect to patient location data: all geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial 3 digits of the zip code if, according to the current publicly available data from the Bureau of the Census: the geographic unit formed by combining all zip codes with the same 3 initial digits contains >20,000 people, and the initial 3 digits of a zip code for all such geographic units containing ≤20,000 people is changed to 000.

An understanding of the HIPAA safe harbor rule has been further muddied by the different ways in which it is described by experts in the fields of public health and geography and by the guidance of the HHS and the OCR. A reader of the *background and context* section on the *2010 De-Identification Standard Workshop* page on the HHS website [11] could justifiably conclude that any aggregation of 20,000 people is in compliance with the safe harbor rule regardless of zip code. In contrast, focusing on the zip code rules as they appear in the literature could lead a person to conclude that zip codes are the primary vehicle for data protection. This is because, in many cases, authors simply do not specify the type of zip code used in their work. This potential for ambiguity among different sources has likely contributed to the number of studies that have aggregated (or suggested the possibility of aggregating) in ways that do not align with the 2000 HIPAA final rule [8,25-27]. Textbox 2 offers a number of different justifications for how scholars have interpreted the safe harbor provisions.

The fact that a range of views exists is not surprising, considering the ways in which HIPAA provisions have been interpreted within the fast-growing scholarly literature using spatial health data and among various web-based help resources. Understanding of the safe harbor provision is muddied by conflicting or ambiguous phrases that appear across a broad array of resources and by how different scholars seem to follow

different practices and procedures for handling patient location data. This profusion of differing practices, although perhaps engendering interesting conversations, likely comes at the cost of research output being unnecessarily overly masked to protect sensitive health data.

Two Different Interpretations

To find a way forward toward more standardized interpretations of HIPAA safe harbor rules, it helps to delineate 2 distinct ways of interpreting the safe harbor provision specific to location data (while recognizing that less common interpretations may also exist). In essence, 2 different and competing interpretations have emerged: the 3-digit zip interpretation and the 5-digit zip interpretation.

The 3-Digit Zip Code Interpretation

For many health researchers, there is only one interpretation of the safe harbor provisions. This is likely because the privacy rule was designed with tabular data in mind, and much medical research involves working with data in its tabular form [9]. For these investigators, a zip code is primarily a 5-digit number that can be reduced to a 3-digit one [5]. For example, an analyst receives a spreadsheet of patient data from which to build a risk model. One column in the table would be designated for the location attribute (ie, a column for zip codes). According to this rule, only the first 3 digits of the zip code are permitted to be shared (unless the population value is <20,000, whereby the data are suppressed or converted to 000). For most lawyers, medical researchers, and those using patient data in tabular format, there is little ambiguity in the safe harbor standard.

The 5-Digit Zip Code Interpretation

For those who view zip code data primarily as spatial data, the privacy rule elicits some confusion. Although a zip code is a 5-digit number, to geographers and a growing number of other scholars who use spatial data, it is also an area on a map. Zip codes divide regions into smaller areas designed to aid post delivery. Both 3-digit zip code areas (Figure 1) and 5-digit zip code areas (Figure 2) are present. The 5-digit zip code areas are nested within 3-digit zip code areas (Figure 3). People who work with spatial data are likely to be familiar with this hierarchy of spatially nesting areas and how it can lead to conflicting interpretations of provision §164.514(b)(2a), which states the following:

> *(2a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people*

In this view, there are 2 ways of reading “Zip codes with the same three initial digits,” namely either (1) using 3-digit zip codes (as described in the previous paragraph) or (2) using 5-digit zip codes that share the same 3 initial digits.

The root of this apparent ambiguity comes from the phrase “all zip codes.” If we interpret “all zip codes” as “all of the five-digit zip codes,” then the 3-digit zip code rule would still apply, as when one combines all the 5-digit zip codes together, they are left with a 3-digit zip code area (Figure 4). However, if “all zip codes” were interpreted as “all five-digit zip codes within the aggregation,” a less conservative interpretation emerges where 5-digit zip codes can be combined to meet the 20,000 population threshold as long as all the used 5-digit zip codes have the same 3 initial digits (Figure 4). Simply put, this interpretation permits investigators to aggregate 5-digit zip codes when they all fall within the same 3-digit zip code area. The large difference in the areas highlighted in Figures 1 and 2 demonstrates the impact of these 2 competing interpretations. Here, we must note that the 5-digit interpretation does not meet HIPAA standards; the reasons for this are discussed later in this paper.

Figure 1. Three-digit zip code boundaries.

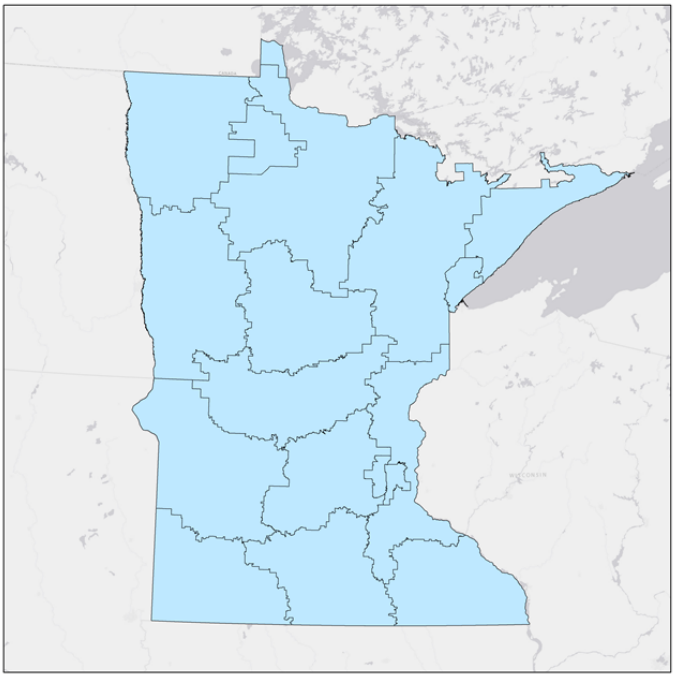


Figure 2. Five-digit zip code boundaries.

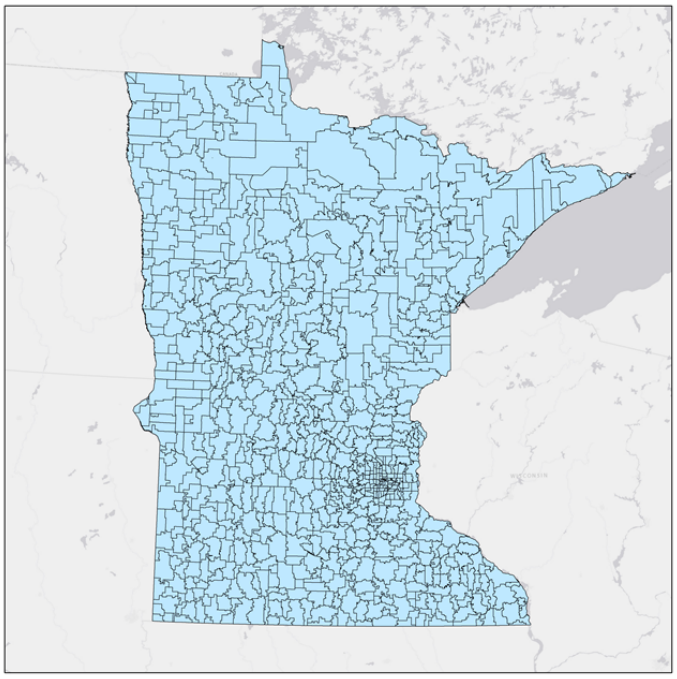


Figure 3. Five-digit zip codes nested within three-digit zip codes.

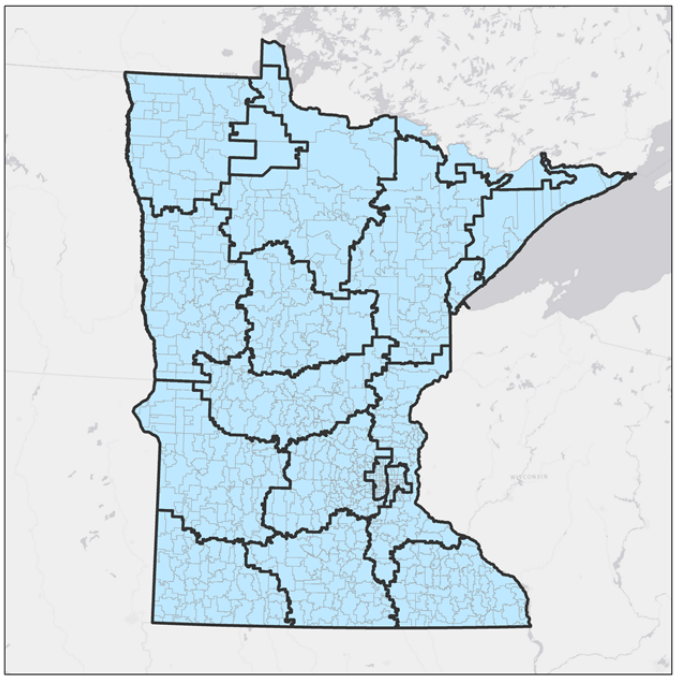
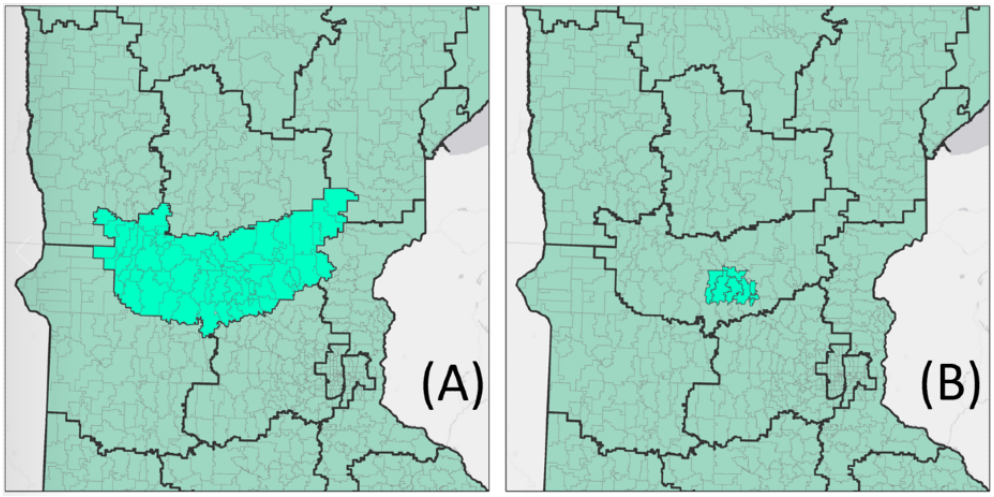


Figure 4. (A) All the 5-digit zip codes beginning in "563." (B) An aggregation of 5-digit zip codes that all begin with "563" and contain >20,000 people.



Drivers and Implications of the 2 Interpretations

Comparing studies that use 3-digit versus 5-digit zip codes illuminates a potential cause for the existence of competing interpretations tied to whether the work uses tabular data or spatial data. In the case of either 3- or 5-digit zip code interpretation, tabular data can appear in essentially the same format (containing only the first 3 digits of a zip code). However, the same mapped data would be *very* different. A researcher operating under the 3-digit interpretation would share maps of patient data at the 3-digit zip code level (Figure 5), and if a 3-digit zip code contained <20,000 people, it would be merged with a neighboring unit. The corresponding tabular data for these maps would only contain 3-digit zip codes. However, investigators operating under the 5-digit zip code interpretation could share maps at the 5-digit zip code level; if the 5-digit zip code contained <20,000 people, it would be merged with neighboring units that share the same first initial digits. The corresponding tabular data for these maps would only contain the first 3 digits of a zip code as well; however, as >1 aggregation would fall within each 3-digit zip code area, there would be multiple records with the same 3-digit zip code.

These differences are not hypothetical as relevant examples are abundant in the literature. Bearing in mind that researchers rarely describe their decision-making in detail, there is a body of work that seems to operate under the 3-digit zip code interpretation [8,10,17,27-30]. Another realm of scholarship appears to operate under the 5-digit zip code interpretation [4,26,30], and there is related work that seems to suggest the capability of aggregating any geocode to meet the 20,000 threshold [7,8,25]. These are some of the many potential examples of how there appears to be a divide between the 3- and 5-digit zip code interpretations of HIPAA.

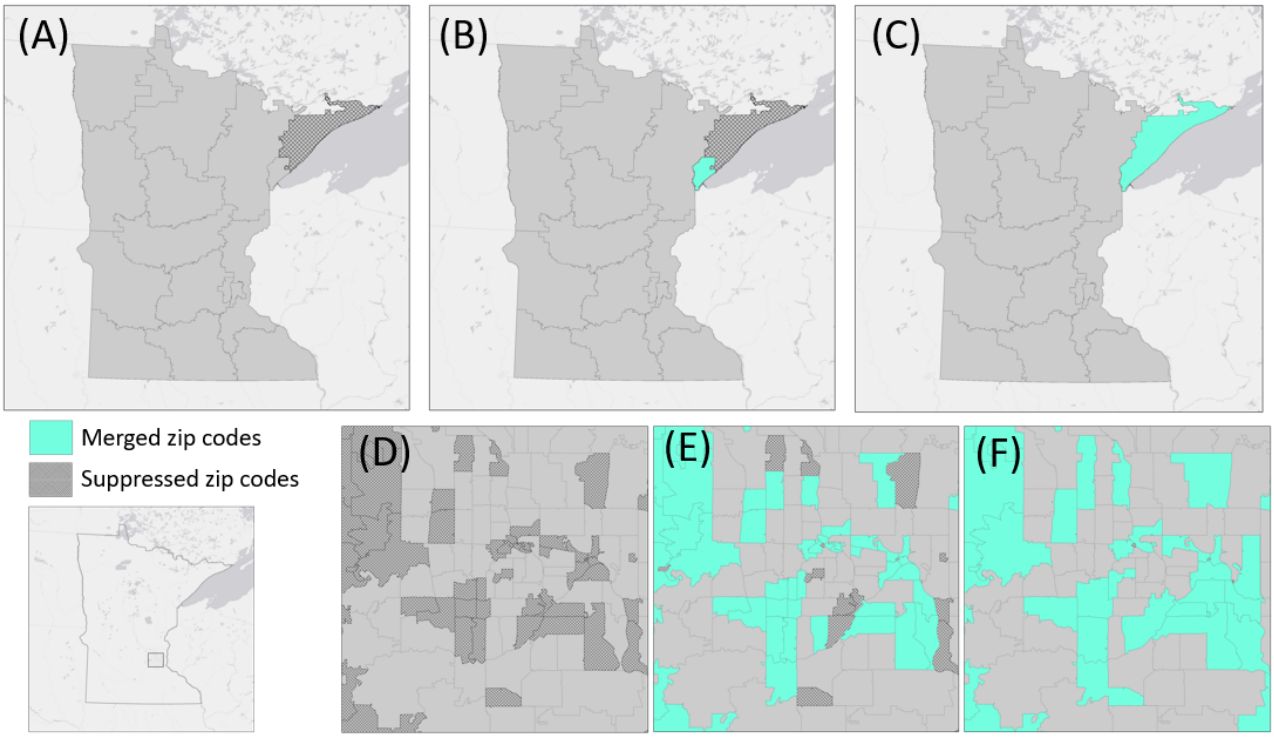
Interestingly, there appears to be some commonality within and differences among disciplines regarding the way a safe harbor is interpreted. Although this paper does not attempt to conduct a full literature review, anecdotally, of the studies cited in the previous paragraph, all those operating under the 3-digit zip code interpretation are authored by epidemiologists, medical researchers, or computer and information scientists, whereas the papers backing the 5-digit zip code interpretation are authored by geographers. Although this is just a sample of a larger literature, there seems to be a trend where spatially oriented researchers are more likely to embrace the 5-digit interpretation or a more lenient understanding of the rules around a threshold of 20,000 people. This is not surprising, given that geographic research often necessitates a map, and 3-digit zip codes are not intuitive map units. It is also the case that 3-digit zip codes are not easy to find in the form of public shapefiles, or mapping files, that are often used for research. Neither Census [31] nor the US Geological Survey offers data at the 3-digit zip code level. In fact, at the time of writing, we can only find 2 sources that provide data for download in the form of 3-digit zip code boundaries for the United States, and both sources are proprietary (Esri's ArcGIS Online and Caliper's Maptitude). Even without access to these proprietary resources, it is possible to create boundaries on one's own. However, one would think that as 3-digit zip codes are the required units for display under HIPAA law, they should be more readily available on the web. In contrast, data at the 5-digit zip code level are easy to find on the web and appear abundantly in the public health literature. The extent to which the dearth of 3-digit zip code map data plays a role in the misunderstanding of the safe harbor rule is unclear; however, one cannot help but wonder whether the widespread confusion would exist if 3-digit zip code mapping files were available for download on the HHS website.

The potential implications of misunderstanding privacy guidelines are profound when considering that researchers share patient data in inconsistent ways that bear on both the efficacy of health interventions and the potential for privacy breaches. When studies share aggregated patient data at the 3-digit zip code level, their output is generally not useful for identifying local distributions of health and disease, although they provide a more generous degree of data security. When studies share PHI at the 5-digit zip code level, they can provide a much more useful depiction of the spatial health dynamics at hand but at the cost of weaker data privacy.

In terms of this trade-off, the difference in identification risk between 3-digit and 5-digit zip codes is substantial enough to warrant an alarm, as discussed in detail in the following section

[15]. At the same time, the difference in spatial resolution between the 2 forms of zip codes carries potentially problematic costs. For instance, one study demonstrated how different disease patterns emerge depending on whether 3-digit or 5-digit zip code areas are used, and with an example data set, the authors showed that if 3-digit zip code areas are used to determine how to best distribute N95 respirators during a pandemic, it would result in a surplus of supplies for health care workers in some communities and shortages others [30].

Figure 5. The aggregation process as seen within (A-C) 3-digit zip codes (D-F) and 5-digit zip codes. Zip codes with populations <20,000 people are suppressed. To address suppression, low-population zip codes are merged with neighboring zip codes to meet Health Insurance Portability and Accountability Act requirements. It is not in adherence with Health Insurance Portability and Accountability Act Safe Harbor to use 5-digit zip codes as the unit of aggregation.



Twin Challenge 2: Data Loss

Overview

Even after gaining a clearer understanding of HIPAA law and how it is meant to be interpreted, one more challenge remains, namely that HIPAA guidelines are very likely too strict in general, resulting in an unnecessarily large degree of data loss [3,17]. The following sections provide insight into the extent of the data loss that occurs when adhering to HIPAA Safe Harbor's 3-digit zip code rule and how other (non-HIPAA–compliant) interpretations can reduce data loss without adding much in terms of privacy risk, depending on the types and amount of data being shared.

Data Loss From 3-Digit Zip Codes and 20,000 People

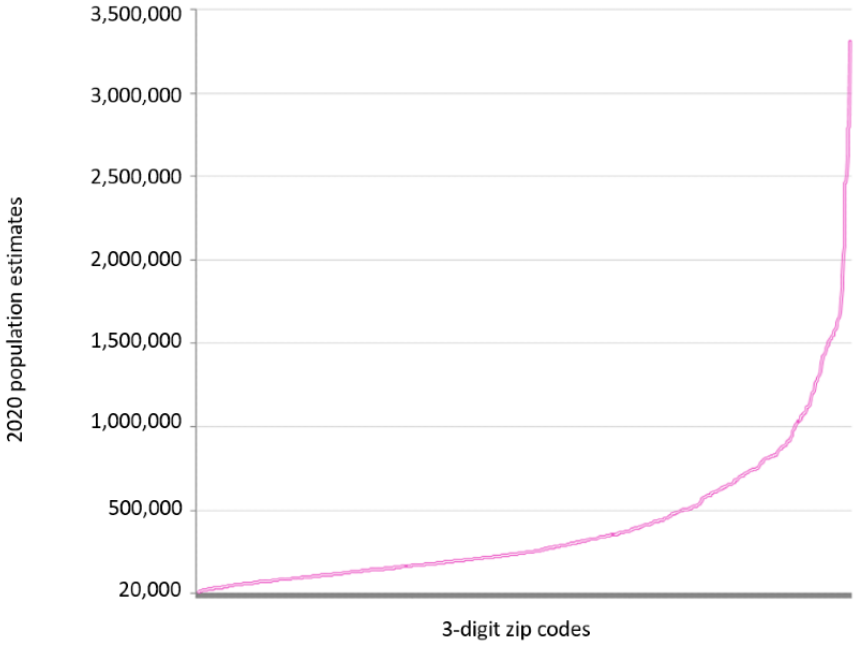
Opting for the 3-digit zip code interpretation is a conservative choice that has a number of negative implications for research and policy. The 3-digit zip code interpretation is very cautious with respect to adhering to the 20,000-person rule. Bear in mind that, as of 2020, the average population contained within a 3-digit zip code is 397,372 people, which is almost 4 times the population threshold of 100,000 required by the Bureau of the Census for the release of microdata (individual response data from the census). Thirty years after the initial rule, there are now only 13 zip codes of 3 digits that require suppression (as they have <20,000 people in them). The number of ideal units containing small but acceptable populations is disappointingly low; only 12 units contain between 20,000 and 30,000 people, and only 21 contain between 30,000 and 40,000 people. Just over 91% of 3-digit zip code geographies contain >60,000 people or at least 3 times the 20,000-person threshold. In simple terms, we should expect that most geographies shared under the 3-digit zip code safe harbor standard will contain populations far greater than the 20,000-person threshold (Figure 6).

Given that most 3-digit zip code geographies contain >20,000 people, under the HIPAA safe harbor provision, most will have a very small proportion of unique records. However, some places will have a proportion of unique records that are considered relatively riskier in terms of patient protection. In any case, the small number of instances that contain the "riskier" low-level minimum populations still meet the minimum acceptable level of risk (which, if we look back at the simulation study by Horm [22], would result in approximately 10% unique records). This is slightly higher than the 7.3% estimated unique records in the 1990 census microdata; however, the HHS points out that the actual risk will be much lower because of the limited number of publicly available tables that can be used to compare the patient data with. These risk estimates are also subject to the myth of the perfect population register, which is discussed later in this paper [17]. Finally, the HHS suggests that the relatively low probability of success should be a deterrent in and of itself.

An interpretation of this threshold is that if the HHS is satisfied with some units being shared at the level of 20,000 people, could all units be shared at that resolution? After all, if populations of 20,000 meet the minimum acceptable level of risk, then what is stopping investigators from aggregating 5-digit zip codes to meet this requirement? Zip codes of 3 digits are rather impractical for research purposes; hence, it is very uncommon to find a map shared at this level. For this reason, it is easy to see how researchers could come to believe that the 5-digit interpretation is permissible if they have not given the legal documents a thorough reading.

Aggregating 5-digit zip codes to create the finest-grained units possible that also still meet the 20,000-person threshold is tempting as this would allow investigators to meet the minimum acceptable level of risk in a way that enables the sharing of maps with more detailed and consistent geographies than that provided by 3-digit zip codes. In this scenario, there would be a slightly greater risk of identification because of the minimum population size, although it would still seem to be an acceptable level of risk as long as the 18 other safe harbor–restricted identifiers were removed. The remaining problem is that 1 of the 18 identifiers is not being *fully* removed in this scenario. By aggregating 5-digit zip codes, an individual record contains more information than a single 3-digit zip code; in addition, it now contains a handful of 5-digit zip codes that can be used to further narrow down the possible matches. Therefore, *5-digit zip code aggregations do not meet HIPAA safe harbor standards*.

Figure 6. Three-digit zip codes (100-999) ordered least to greatest by population from 2020 estimates from the American Community Survey.



However, depending on what other information is kept, it is reasonable to believe that sharing a map of patient data stripped of age and other demographics at the aggregated 5-digit zip code level would lead to a very low (certainly quite low) risk of identification. One study showed that certain elements from a list of 18 identifiers can still be shared without jeopardizing patient privacy "when other features are reduced in granularity." Specifically, Malin et al [28] found that more detailed age data (beyond what is permitted by safe harbors) could be shared when they coarsened the specificity of other variables such as ethnicity [28]. The authors noted that every data set is different, and because of this, alternative deidentification practices can be used to enable the safe disclosure of patient data that are normally suppressed under the safe harbor method. This means that there is potential for 5-digit zip code information to be safely shared in an aggregated form as long as other identifying information is suppressed.

In summary, it may be time to rethink the one-size-fits-all strategy, which is the safe harbor method. It is reasonable to ask whether aggregating 5-digit zip codes into regions that contain at least 20,000 people could achieve a "sufficiently low" risk of identification when other patient information is suppressed, such as date of birth (DoB) and gender. It would be even more reasonable to suggest that aggregating 5-digit zip codes could work if no patient information other than diagnosis and location was shared. Curtis et al [6] tested this claim in a study that found that when put to the test, students were unable to identify individuals in simulated cancer maps. There was little reengineering risk, even at aggregated resolutions of finer than 20,000 people. To this point, this paper has pointed out the ambiguities within the safe harbor standard while shedding light on some of the arbitrary determinations made by the HHS that have contributed to a perhaps overly conservative definition of privacy. The following section takes a closer look at how the safe harbor rule has been criticized for being too stringent and, at the same time, not sufficiently protective, specifically when it comes to identification risk.

Do the Privacy Gains Justify the Amount of Data Loss?

To dive deeper, we must go back and consider the influence of the population-level identification attack by Sweeney [15]. As stated previously, this initially resulted in the decision to bar both 3-digit and 5-digit zip codes from deidentified data; however, after taking public comments, the HHS reconsidered,

and 3-digit zip codes were deemed permissible as long as they contained a population of at least 20,000 people. The HHS justified their restrictions by citing particular studies that led them to believe that the combination of 5-digit zip code, gender, and DoB would be enough to potentially identify a great deal (more than half) of the US population based on uniqueness [32]. Note that to be considered "unique," a record must contain a combination of characteristics that make it different from all other records in that table [33]. If the number of unique individuals within the US population was as large as Sweeney [15] reported, the motion to block the 5-digit zip code and DoB under safe harbor seems quite justified. However, some have pointed out that the combination of these 3 identifiers, even with their formidable discernibility capabilities, might not be as threatening as the article by Sweeney [15] makes it out to be.

Barth-Jones [17] describes the "myth of the perfect population register" in his 2012 paper, which points out how many investigators often forget to account for the people missing from the lists used to link individuals to their medical records. These missing populations add significant uncertainty to the calculation of true population uniqueness [17]. Therefore, the actual proportion of unique individuals on a list cannot be determined with 100% certainty if potential matches exist off the list. Therefore, these kinds of studies must be careful in the statements they make—oftentimes including phrases such as "likely unique" or "potentially identifying" as certain identification cannot be claimed without a list of the entire population or the knowledge that the individual under identification attack was indeed contained within both lists.

For instance, consider the paper by Sweeney [15], which the 1999 NPRM cites saying "A 1997 MIT study showed that because of the public availability of the Cambridge, Massachusetts voting list, 97 percent of the individuals in Cambridge whose data appeared in a data base which contained only their nine-digit zip code and DoB could be identified with certainty." [16] According to this, nearly all Cambridge voters can be identified using the combination of DoB and 9-digit zip code. Sweeney [15] states that this proportion of people can be "uniquely identified" on this basis; however, these individuals are only uniquely identifiable within the population of registered voters and not within the general Cambridge population (see the study by Barth-Jones [17] for a full explanation). This means that, for an intruder to identify an individual's medical record, they must know that the individual exists on both lists and that no other person in Cambridge shares the same DoB and 9-digit zip code. When deciphering the data, the intruder must account for 35,000 nonregistered voting-aged people living in the city, any of whom could be the true subject of the medical record of interest. Unaccounted populations inject much uncertainty into the identification of unique records (35% error in the study by Sweeney [15]). With an imperfect population register, as exemplified by the Cambridge attack, an intruder would be able to identify no one with 100% certainty. Barth-Jones [17] concludes that the governor was likely only identifiable based on the fact that he was a public figure who had public hospitalization. The date of hospitalization was known, as well as his DoB, gender, and zip code; moreover, it could be easily assumed that he would be a registered voter. In instances such as this (having information a priori), an intruder can be confident of a unique match.

It is unclear whether the HHS wrote the NPRM with a full understanding of the methodological limitations of voter list-based identity attacks of the kind described by Barth-Jones [17]. It is possible that the clause "...could be identified with certainty" was taken without really considering the implications of the prior clause "...whose data appeared in the data base." Many assumptions must be met before we can ignore the myth of the perfect population register. In this example, to identify 97% of the individuals with certainty, we would need to be sure that none of the 54,805 voters on the voter list had the same birth date as a nonvoter living in their neighborhood. We might then wonder how 97% could be identified on the list compared with the proportion identifiable in the entire Cambridge population. This is something we cannot determine as we do not have a population register. However, given that the total population of Cambridge is approximately 88,000 [17], there is much room for error. If the HHS based its development of safe harbors on a limited understanding of these complexities, it might lead us to wonder whether the level of protection delineated within the safe harbor standard is overly conservative.

Nevertheless, even if the HHS misunderstood how Sweeney [15] was using the term "identifiable" in her 1997 paper, there is still room for concern about how far to read into the study. The work by Sweeney [15] is bold, insightful, and conveys a critical message: private information is vulnerable to attacks. The extent to which we *understand* the vulnerability is unclear. Even with the injection of uncertainty from missing populations, the risk for identification may still be considered too high and the implications would be quite serious. Let us return to the Barth-Jones [17] review of the attack by Sweeney [15], which finds that somewhat fewer (but perhaps not much fewer) than 29,000 people out of 88,000 in Cambridge are identifiable (if the record is unique and the data intruder already knows that the individual is on both lists). Depending on the motive of the data intruder, this might not be far from likely. It is easier to link a specific person to their medical record than to link a specific medical record to the person to which it belongs. This is because a motivated attacker is likely to have collected background information on the person a priori. The data intruder likely has a target in mind—someone they know—and therefore, it is not that unlikely for them to already have information on the target's voting behaviors and place of work, allowing the intruder to determine the employment insurance coverage that could be used to confirm the target's presence on the insurance hospitalization data list. Moreover, even without knowing with certainty if the target of the attack is on both lists, the fact that the chance of a false positive (matching a record to a voter on the list when the record actually belongs to a nonregistered voter) occurring could be perceived as highly unlikely by the attacker, which could encourage them to continue with their plans regardless of the potential false positive.

The combination of DoB, gender, and 5-digit zip codes can be problematic when shared in conjunction. The question that remains is whether this combination of identifiers can be reworked to reduce the risk of identification. In the literature

on microdata anonymity, zip code, gender, and DoB are actually not considered full identifiers themselves but rather quasi-identifiers that can be used in combination to find unique instances. The term "identifier" is reserved for information that uniquely identifies an individual, such as a social security number [34]. Nevertheless, quasi-identifiers can be dangerous when used in combination; however, how dangerous are they? To gain some insight into this question, we must look more closely at how identification risk has appeared in the literature, relying on the HIPAA safe harbor method.

What Level of Data Loss Defines Sufficient Data Protection?

What is the acceptable level of identification risk? There is no universally recognized standard that defines what a sufficient proportion of unique records should be. Some have suggested that the nationally accepted standard of reidentification risk is defined by HIPAA's safe harbor standard itself [27] but recall that the safe harbor standard was derived somewhat arbitrarily, being loosely based on rules used by the Bureau of the Census and a couple of simulation studies. In fact, when determining the population requirement of the HIPAA safe harbor rule, the HHS made the following statement in regard to defining "minimal risk":

> *With respect to how we might clarify the requirement to achieve a "low probability" that information could be identified, the Statistical Policy Working Paper 22 referenced [see 18 in our references] discusses the attempts of several researchers to define mathematical measures of disclosure risk only to conclude that "more research into defining a computable measure of risk is necessary." When we considered whether we could specify a maximum level of risk of disclosure with some precision (such as a probability or risk of identification of <0.01), we concluded that it is premature to assign mathematical precision to the "art" of deidentification.*

Twenty years later, there is still no threshold defining "sufficiently low probability," and investigators fall back on the safe harbor standard as a point of reference for comparing different levels of data protection. Deidentification with the safe harbor method is said to leave somewhere around 0.03% or 0.04% of records within the US population vulnerable to identification [17,35]; however, this proportion fluctuates according to the geographical extent of the data set, where some regions have much smaller proportions of unique records and others have much higher. Specifically, the reidentification risk has been found to range from 0.01% to 0.19% [28], 0.01% to 0.25% [36], and 0.013% to 0.22% [37] on a state-by-state basis.

Most studies estimate the identification risk under a safe harbor to be low. However, there is no consensus on whether safe harbor standards are sufficient to protect patient data. In other words, "sufficiently de-identified" is subjective and, on occasion, very similar proportions of unique records have evoked very different assessments. For example, Sweeney asserts that the estimated safe harbor reidentification risk of 0.04% of the US population is not a sufficient privacy guard [10,35], whereas Barth-Jones [17] suggested that the risk would actually be <0.03% (when using a voter list attack strategy) and that this proportion is, in fact, sufficient; he goes on to compare the identification risk under a safe harbor to the likelihood of being struck by lightning [17]. A reidentification attack by Kwok et al [37] reidentified only 2 of 15,000 individuals (0.013%) from a safe harbor protected data set, and the intruder was provided with a substantial amount of information from a market research company. Kwok et al [37] concluded that there was a low risk of reidentification and that masking with a safe harbor makes reidentification a challenging task. Others asserted that the safe harbor is too stringent. Malin et al [28] suggested in a 2011 article that the safe harbor method was too conservative as it is possible to release more detailed information without presenting a greater risk than that provided by the safe harbor method. In contrast, a 2016 study found that even when data seem sufficiently masked, computer science models can be used to identify a large proportion (42.8%) of patients by linking demographics such as age, sex, hospital, and year [38]. Although specific to a single case study, this is a high and likely unacceptable level of risk. More recently, Janmey and Elkin [27] suggested that the safe harbor standard is sufficient for preserving privacy at an overall population level. However, they also found that encounter notes within data can sometimes include indirect identifiers that can be used to help match records, and this could increase the risk of identification to 0.07%, which is well over the estimated range of risk previously mentioned when using safe harbor [17,35].

It is safe to say that there is disagreement regarding what is sufficient for data protection. This type of risk calculation is complicated in and of itself and a concept such as *sufficiency* is necessarily a judgment call. Identification risk depends not only on how the data are released but also on the alternative lists publicly available to the data intruder. Sweeney [10] described how identification risk for safe harbor–abiding data sets can be as high as 25% when the intruder uses more than just a voter registration list. Other detailed registries can be used to reidentify masked data such as real estate tax data, credit reports, and property records. Moreover, identification risk can foreseeably jump much higher—far beyond the expected ranges—for certain areas where the demographics of the base populations allow an intruder to easily narrow down potential matches based on age or ethnicity, as seen in regions dominated by college dorms, ethnic enclaves, or transient communities [15,38]. Sufficient data protection (leaving aside the definition of sufficiency) will always be dependent on the data set being masked as a slew of factors determines the overall identification risk.

Ways Forward

Overview

So far, we have focused on 2 key issues of safe harbor provisions: the confusion around which zip codes to use and whether the rule warrants an unnecessarily large amount of data loss. Reviewing the process by which the safe harbor concept came into being provides insight into the intended interpretation of the provision and the motivations that guided its development; however, this is the first step. The ambiguity of how to best

interpret and use zip codes or other geographic identifiers persists, and there is no clear consensus on what defines sufficient minimal risk. In this paper, we explore new approaches to data privacy and how they may meet the needs of some researchers; however, we conclude by arguing that the most promising way forward to addressing the twin problems of safe harbor is to steer away from one-size-fits-all guidelines and toward deeper assessments of domain-specific and data-specific modes of masking that could offer a middle ground between useful data and protected data.

New Approaches to Deidentification

In the face of the complex nature of reidentification risk, scholars and policy makers have begun to advocate for the widespread adoption of k-anonymity or differential privacy (DP) methods [10]. The primary argument for these approaches is that deidentification methods should come with privacy guarantees, especially as technology advances and powerful automated systems can be made to search for matches between multiple public lists. Therefore, although k-anonymity and DP cannot necessarily guarantee data security, these methods have been receiving considerable attention recently as they provide a type of privacy guarantee that offers more complete data protection than traditional masking approaches.

K-anonymity ensures that no unique records exist in the data set and further requires that each record has a minimum of "k-1" common records (those that have the same quasi-identifiers) so that they cannot be differentiated and therefore identified with certainty [39]. K-anonymity can be achieved through many traditional methods such as jittering, aggregation, and location swapping, and it often provides a higher level of protection than if one were to use one of these traditional methods alone. However, k-anonymity is not impervious to intruder attacks. An intruder can still use background knowledge to narrow down the possible matches to increase the likelihood of identification, such as in a homogeneity attack (attacks based on data that contain identical values for an attribute), in which a region with a homogeneous population containing similar values for a record in the table can be used (alone or linked with other data) to identify an individual or diagnosis. Therefore, k-anonymity, strictly speaking, does not guarantee privacy. However, it guarantees nonuniqueness, which, in the absence of outside knowledge, provides considerable data protection, and therefore, k-anonymity remains a popular approach.

DP is attracting attention as a newer approach to protecting sensitive data that assures a very low likelihood of individual identification. The most common definition of DP is that of epsilon DP introduced by Dwork et al [40]. The epsilon DP by Dwork et al [40] involves creating a synthetic aggregated data set from an original unprotected data set, which ensures that an individual record cannot be identified. These simulated data are built by injecting a predetermined amount of noise (based on a Laplace distribution) into the original aggregate table such that it does not significantly influence the output (of queries into particular prespecified relationships). In other words, the aggregate table is systematically adjusted to secure individual privacy while also ensuring that the data provide similar results to what would have been given if the original data were used in a prespecified analytical model. This is achieved such that if any one individual was removed from the data set, it would not influence the overall results. This means that epsilon DP provides relative guarantees about disclosure risk, and essentially promises that "...any given disclosure will be, within a small multiplicative factor, just as likely whether or not the individual participates in the database." [40]

Unlike k-anonymity, DP protects data under the assumption that an intruder has close to perfect knowledge, and in doing so, DP offers a level of protection unlike others. DP does not succumb to the same weaknesses as traditional methods (including the homogeneity attack) and provides stronger data protection against differencing, linkage, and reconstruction attacks [41]. In addition, because of its robustness, DP has the advantage of reducing improper data analysis techniques by limiting the ability of a single observation to have an effect on the result, which helps to deter things such as p-hacking, hypothesizing after the results are known, and overfitting models [42]. For these, and many other reasons, DP has gained considerable attention over the past 2 decades. In fact, DP methods have the potential to replace existing masking methods and have already been adopted by Apple and the Bureau of the Census, which intends to use DP to protect the 2020 census microdata. DP is not infallible; it offers "an extremely strong guarantee, it does not promise unconditional freedom from harm." [41]

As DP provides a higher level of protection than many other methods, it potentially offers a way for researchers to share data at more detailed levels than previously allowed in safe harbors. In an example of disease surveillance mapping, the safe harbors' minimum population requirement of 20,000 people is rather limited in terms of map resolution. A map with units containing 20,000 people would not provide enough detail to be helpful to researchers, policymakers, or community members. However, DP would allow investigators to share maps at much finer scales (down to the neighborhood level) without putting patient identities at risk.

Thus, why not use DP? This is because it has critical drawbacks for research use [43]. For instance, a map created from a differentially private aggregated table displays simulated data; therefore, it is possible that some regions on the map would not accurately reflect the original data, especially at finer scales where the population numbers are lower. Santos-Lozada et al [44] found that the infusion of noise from DP methods affects observed distributions differently for different demographics, meaning that DP has the potential to bias the understanding of health disparities at the national level. In particular, the authors demonstrated how mapping differentially private data led to "overestimates of population-level health metrics of minority populations in smaller areas and underestimates of mortality levels in more populated ones," and these effects were dramatic. For instance, note the following:

> *...in McCulloch County, Texas, the mortality rate ratio for non-Hispanic blacks is 75.9, indicating the mortality rate would be 24% lower under the current methodology compared with the differential privacy methodology. Similarly, in Clarke County, Virginia,*

> *the mortality rate ratio for Hispanics is 121.4, indicating the mortality rate would be 21% higher under the current methodology compared with the differential privacy methodology. At the same time, the non-Hispanic white mortality rate ratios were essentially unchanged for these two counties, at 100.3 and 99.8, respectively, meaning substantial biases may enter into understandings of disparities.*

The implications of DP for research are dire, and the recent move by the Bureau of the Census to adopt this approach for the 2020 census microdata has drawn much attention to its advantages and disadvantages [45,46]. Census data are one of the largest sources of sociodemographic data used by social scientists; therefore, differentially private methods threaten to degrade the reliability and effectiveness of social science research. Other than threats to data accuracy and biases, another source of concern regarding the 2020 census data is that these differentially private tables would not enable exploratory data analysis. This is because differentially private data are synthetic, and therefore, relationships cannot be explored unless they are prespecified when the synthetic table is created. For this, it is very likely for DP to interfere with the process of data-driven scientific research, pushing some scholars to suggest that perhaps "...differential privacy goes far beyond what is necessary to keep data safe" [46].

There is much uncertainty regarding the practicality of DP for the protection of large-scale, sensitive data. DP is a relatively new concept for several social scientists and epidemiologists. There is a dearth of investigations into DP within the social science literature, particularly regarding the impact it might have on health mapping. We could only find 1 study at the time of writing [44] but expected more, given the attention paid to DP and the many unanswered questions that it poses. What are the implications of DP in mapping in terms of accuracy and use? How do differentially private maps compare with maps of the original raw data? Furthermore, it is unclear how DP stands within institutional review boards. This is relatively new territory, and it is likely that many HIPAA compliance officers are unfamiliar with DP. As part of our examination of the history of HIPAA, we spoke with legal experts and HIPAA compliance officers. One such officer, on being introduced to DP, stated that "this doesn't play into our office's considerations of deidentification." DP holds some promise for mapping spatial data but at known and unknown costs.

Current State and Future Research

Despite the ongoing interest in expanding the use and sharing of health data mapping, the safe harbor rule stands as the primary guidance for those interested in sharing maps. It is far from perfect in that for many scholars, it is ambiguous and either too stringent or insufficient in terms of securing data or reducing data loss. Alternative methods exist, which have the potential to do a better job; however, they have their own drawbacks. HIPAA safe harbor provisions do not set out to guarantee data protection similar to the newer modes of data protection; instead, they only ensure a low risk of identification with the ultimate goal being "to balance the needs of the individual with the needs of the society" [18]. The challenge is to find the "sweet spot" between protected data and useful data while also understanding that this sweet spot changes for each data set depending on what and how much information is available to the public. Furthermore, with rapidly evolving technologies, this sweet spot will continue to change over time. The amount of individual-level data collected by companies today is large and continuously growing. In fact, society may have already reached the point where the myth of the perfect population register is no longer a myth in the face of big data [47].

Although safe harbor continues to stand as the primary source of guidance for handling spatial health data, researchers continue to work with and against it in ways that reflect their understanding of the law and their data against a larger sociotechnical backdrop. As demonstrated by Malin et al [28], there are ways of safely sharing more detailed data (ie, age information) by coarsening the granularity of other data. From this example, we can assume that there are also ways of sharing fine-grained geographic data by censoring other elements in the data. Given that some pieces of information contribute more heavily to individual identification than others (ie, DoB being more identifying than gender), we are left to ask questions that, if answered, could help inform future approaches. Could a 5-digit zip code become innocuous without age information? How many individuals can be uniquely identified by age and 5-digit zip code alone? What if all age and gender information were removed? Would a 5-digit zip code still have the power to identify an individual? In other words, is it reckless to share maps at the 5-digit zip code level if all other patient information is removed (ie, only the sharing of the 5-digit zip code and diagnosis)? What if these zip codes were aggregated to form units that each contained 20,000 people within them? What would be the risk for identification? Of course, it is easier to ask these questions than answer them; however, by examining the history of HIPAA and clarifying the importance of 3-digit zip codes versus 5-digit zip codes, we have a stronger foundation for answering these questions. Until then, the safe harbor method stands as our primary mode of guidance, and 2 decades after its introduction, these guidelines do not meet the public's need for data security or researchers' need for useful data.

Conclusions

Vague privacy provisions stand as an obstacle to progress and pose a threat to public privacy by hindering the ways in which epidemiologists and geographers understand how to share spatial data. This paper promotes an understanding of the HIPAA safe harbor provision by providing a comprehensive overview of the law while also presenting various expert perspectives and relevant studies that, taken together, show how alternative methods to safe harbor can offer researchers better data and better data protection. Two different interpretations of the safe harbor rule exist—the 3-digit and 5-digit zip code interpretation—and although 5-digit zip codes are not the intended level of aggregation under the rule, there is reason to believe that information can be safely shared on a map at this level. More research is needed to determine whether the risk for individual identification is sufficiently low for maps shared at the 5-digit zip code level when DoB and gender are suppressed from the map's corresponding table. Much has changed in the 20 years since the introduction of the safe harbor

provision; however, it continues to be the primary source of guidance (and frustration) for researchers trying to share maps, leaving many waiting for these rules to be revised in accordance with the times.

Availability of Data and Material

Data sharing is not applicable to this paper as no data sets were generated or analyzed during the study.

Authors' Contributions

BK drafted the first version of the manuscript. BK was responsible for data acquisition, data analysis, and interpretation. BK and SMM edited and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. Guidance on De-identification of Protected Health Information. 2012. URL: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf [accessed 2022-06-22]
2. Gupta A, Lai A, Mozersky J, Ma X, Walsh H, DuBois JM. Enabling qualitative research data sharing using a natural language processing pipeline for deidentification: moving beyond HIPAA Safe Harbor identifiers. JAMIA Open 2021 Jul;4(3):ooab069 [FREE Full text] [doi: 10.1093/jamiaopen/ooab069] [Medline: 34435175]
3. Clause SL, Triller DM, Bornhorst CP, Hamilton RA, Cosler LE. Conforming to HIPAA regulations and compilation of research data. Am J Health Syst Pharm 2004 May 15;61(10):1025-1031. [doi: 10.1093/ajhp/61.10.1025] [Medline: 15160778]
4. Curtis A. From healthy start to hurricane Katrina: using GIS to eliminate disparities in perinatal health. Stat Med 2008 Sep 10;27(20):3984-3997. [doi: 10.1002/sim.3260] [Medline: 18381702]
5. Paul O. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev 2009;57:1701.
6. Curtis A, Mills JW, Agustin L, Cockburn M. Confidentiality risks in fine scale aggregations of health data. Comput Environ Urban Syst 2011 Jan;35(1):57-64. [doi: 10.1016/j.compenvurbsys.2010.08.002]
7. Jung H, El Emam K. A linear programming model for preserving privacy when disclosing patient spatial information for secondary purposes. Int J Health Geogr 2014;13(1):16. [doi: 10.1186/1476-072x-13-16]
8. Browne AC, Kayaalp M, Dodd ZA, Sagan P, McDonald CJ. The challenges of creating a gold standard for de-identification research. AMIA Annu Symp Proc 2014;2014:353-358 [FREE Full text] [Medline: 25954338]
9. Kayaalp M, Browne AC, Sagan P, McGee T, McDonald CJ. Challenges and insights in using HIPAA privacy rule for clinical text annotation. AMIA Annu Symp Proc 2015;2015:707-716 [FREE Full text] [Medline: 26958206]
10. Sweeney L, Yoo J, Perovich L, Boronow K, Brown P, Brody J. Re-identification Risks in HIPAA Safe Harbor Data: a study of data from one environmental health study. Technol Sci 2017;2017:2017082801. [Medline: 30687852]
11. Workshop on the HIPAA privacy rule's de-identification standard. HHS.gov. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/2010-de-identification-workshop/index.html [accessed 2021-01-15]
12. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2018 Mar 06;25(S 01):S48-S61. [doi: 10.15265/iys-2016-s006]
13. Best SJ, Krueger BS, Ladewig J. Privacy in the information age. Public Opinion Q 2006 Aug 25;70(3):375-401. [doi: 10.1093/poq/nfl018]
14. All Information (Except Text) for H.R.3845 - District of Columbia Appropriations Act, 1997. CONGRESS.GOV. URL: https://www.congress.gov/bill/104th-congress/house-bill/3845/all-info [accessed 2021-01-15]
15. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly System. Proc AMIA Annu Fall Symp 1997:51-55 [FREE Full text] [Medline: 9357587]
16. Standards for privacy of individually identifiable health information. In: The Privacy Papers. Boca Raton, Florida: Auerbach Publications; 2001.
17. Barth-Jones DC. The 're-identification' of governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. SSRN J 2012. [doi: 10.2139/ssrn.2076397]
18. Standards for privacy of individually identifiable health information. In: The Privacy Papers. Boca Raton, Fla: Auerbach Publications; 2001.
19. Standards for Privacy of Individually Identifiable Health Information. In: The Privacy Papers. Boca Raton, Florida: Auerbach Publications; 2001.
20. Statistical Policy Working Paper 22. The Federal Committee on Statistical Methodology (FCSM). 1994. URL: https://nces.ed.gov/FCSM/pdf/spwp22.pdf [accessed 2022-07-12]

21. The geographic component of disclosure risk for microdata. United States Census Bureau. Jul 7. URL: https://www.census.gov/library/working-papers/1990/adrm/rr90-13.html [accessed 2022-07-24]
22. 22 HJ. A simulation study of the identifiability of survey respondents when their community of residence is known. National Center for Health Statistics 2000.
23. Department of health and human services 45 CFR Parts 160 and 164 Standards for privacy of individually identifiable health information; final rule. Federal Register. 2000. URL: https://www.govinfo.gov/app/details/FR-2000-12-28/00-32678 [accessed 2022-07-27]
24. ESRI homepage. ESRI. URL: http://www.esri.com/software/businessanalyst) [accessed 2022-06-21]
25. Mu L, Wang F, Chen VW, Wu X. A place-oriented, mixed-level regionalization method for constructing geographic areas in health data dissemination and analysis. Ann Assoc Am Geogr 2014 Dec 10;105(1):48-66 [FREE Full text] [doi: 10.1080/00045608.2014.968910] [Medline: 26251551]
26. Acevedo-Garcia D. Zip code-level risk factors for tuberculosis: neighborhood environment and residential segregation in New Jersey, 1985-1992. Am J Public Health 2001 May 01;91(5):734-741. [doi: 10.2105/ajph.91.5.734] [Medline: 11344881]
27. Janmey V, Elkin PL. Re-identification risk in HIPAA de-identified datasets: the MVA Attack. AMIA Annu Symp Proc 2018;2018:1329-1337 [FREE Full text] [Medline: 30815177]
28. Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. J Am Med Inform Assoc 2011 Jan 01;18(1):3-10 [FREE Full text] [doi: 10.1136/jamia.2010.004622] [Medline: 21169618]
29. Nicholson S, Smith CA. Using lessons from health care to protect the privacy of library users: guidelines for the de-identification of library data based on HIPAA. J Am Soc Inf Sci 2007 Jun;58(8):1198-1206. [doi: 10.1002/asi.20600]
30. Tellman N, Litt ER, Knapp C, Eagan A, Cheng J, Radonovich LJJ. The effects of the Health Insurance Portability and Accountability Act privacy rule on influenza research using geographical information systems. Geospat Health 2010 Nov 01;5(1):3-9. [doi: 10.4081/gh.2010.182] [Medline: 21080316]
31. Nation continues to age as it becomes more diverse. United States Census Bureau. URL: https://www.census.gov/ [accessed 2022-07-25]
32. Simple demographics often identify people uniquely. Carnegie Mellon University. URL: https://dataprivacylab.org/projects/identifiability/paper1.pdf [accessed 2022-07-24]
33. Estimation of the number of unique population elements using a sample. Bureau of the Census. URL: http://www.asasrms.org/Proceedings/papers/1991_061.pdf [accessed 2022-07-24]
34. Microdata protection. In: Secure Data Management in Decentralized Systems. Boston, MA: Springer; 2007.
35. Enhanced protections for uses of health data: a stewardship framework for “secondary uses” of electronically collected and transmitted health data. Secretary of the U.S. Department of Health and Human Services. URL: https://tinyurl.com/3dptn9rh [accessed 2021-01-15]
36. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc 2010;17(2):169-177 [FREE Full text] [doi: 10.1136/jamia.2009.000026] [Medline: 20190059]
37. Kwok P, Davern M, Hair E, Lafky D. Harder than you think: a case study of re-identification risk of HIPAA-compliant records. In: Proceedings of the 2011 Joint Statistical Meetings. 2011 Presented at: 2011 Joint Statistical Meetings; Aug 2, 2011; Chicago, IL.
38. O’Neill L, Dexter F, Zhang N. The risks to patient privacy from publishing data from clinical anesthesia studies. Anesthesia Analgesia 2016;122(6):2017-2027. [doi: 10.1213/ane.0000000000001331]
39. Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information (abstract). In: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. 1998 Presented at: SIGMOD/PODS98: Special Interest Group on Management of Data; Jun 1 - 4, 1998; Seattle Washington USA. [doi: 10.1145/275487.275508]
40. Differential privacy. In: Automata, Languages and Programming. Berlin, Heidelberg: Springer; 2006.
41. Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT Theoretical Comput Sci 2014;9(3-4):211-407. [doi: 10.1561/0400000042]
42. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth AR. Preserving statistical validity in adaptive data analysis. In: Proceedings of the forty-seventh annual ACM symposium on Theory of Computing. 2015 Presented at: STOC '15: Symposium on Theory of Computing; Jun 14 - 17, 2015; Portland Oregon USA. [doi: 10.1145/2746539.2746580]
43. Muralidhar K, Domingo-Ferrer J, Martínez S. -Differential privacy for microdata releases does not guarantee confidentiality (let alone utility). In: Privacy in Statistical Databases. Cham: Springer; 2020.
44. Santos-Lozada AR, Howard JT, Verdery AM. How differential privacy will affect our understanding of health disparities in the United States. Proc Natl Acad Sci U S A 2020 Jun 16;117(24):13405-13412 [FREE Full text] [doi: 10.1073/pnas.2003714117] [Medline: 32467167]
45. Oberski DL, Kreuter F. Differential privacy and social science: an urgent puzzle. Harvard Data Sci Rev 2020 Jan 31;2(1). [doi: 10.1162/99608f92.63a22079]
46. Ruggles S, Fitch C, Magnuson D, Schroeder J. Differential privacy and census data: implications for social and economic research. AEA Papers Proceedings 2019 May 01;109:403-408. [doi: 10.1257/pandp.20191107]

47. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy (sp 2008). 2008 Presented at: 2008 IEEE Symposium on Security and Privacy (sp 2008); May 18-22, 2008; Oakland, CA, USA. [doi: 10.1109/sp.2008.33]

Abbreviations

DoB: date of birth
DP: differential privacy
HHS: Department of Health and Human Services
HIPAA: Health Insurance Portability and Accountability Act
NPRM: Notice of Proposed Rulemaking
OCR: Office of Civil Rights
PHI: protected health information

Edited by C Lovis; submitted 05.03.22; peer-reviewed by L Nweke, D Reuter, J Ropero; comments to author 02.06.22; revised version received 23.06.22; accepted 27.06.22; published 03.08.22.

Please cite as:
Krzyzanowski B, Manson SM
Twenty Years of the Health Insurance Portability and Accountability Act Safe Harbor Provision: Unsolved Challenges and Ways Forward
JMIR Med Inform 2022;10(8):e37756
URL: https://medinform.jmir.org/2022/8/e37756
doi:10.2196/37756
PMID:35921140

Viewpoint

Tempering Expectations on the Medical Artificial Intelligence Revolution: The Medical Trainee Viewpoint

Zoe Hu[1*], BSc, MD; Ricky Hu[1,2*], BASc, MASc; Olivia Yau[3], BSc; Minnie Teng[3], MSc; Patrick Wang[1], BHSc, MD; Grace Hu[4], BSc; Rohit Singla[2,3], BASc, MASc

[1]School of Medicine, Queen's University, Kingston, ON, Canada

[2]School of Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada

[3]School of Medicine, University of British Columbia, Vancouver, BC, Canada

[4]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Zoe Hu, BSc, MD
School of Medicine
Queen's University
166 Brock Street
Kingston, ON, K7L5G2
Canada
Phone: 1 6132042952
Email: zhu@qmed.ca

Abstract

The rapid development of artificial intelligence (AI) in medicine has resulted in an increased number of applications deployed in clinical trials. AI tools have been developed with goals of improving diagnostic accuracy, workflow efficiency through automation, and discovery of novel features in clinical data. There is subsequent concern on the role of AI in replacing existing tasks traditionally entrusted to physicians. This has implications for medical trainees who may make decisions based on the perception of how disruptive AI may be to their future career. This commentary discusses current barriers to AI adoption to moderate concerns of the role of AI in the clinical setting, particularly as a standalone tool that replaces physicians. Technical limitations of AI include generalizability of performance and deficits in existing infrastructure to accommodate data, both of which are less obvious in pilot studies, where high performance is achieved in a controlled data processing environment. Economic limitations include rigorous regulatory requirements to deploy medical devices safely, particularly if AI is to replace human decision-making. Ethical guidelines are also required in the event of dysfunction to identify responsibility of the developer of the tool, health care authority, and patient. The consequences are apparent when identifying the scope of existing AI tools, most of which aim to be physician assisting rather than a physician replacement. The combination of the limitations will delay the onset of ubiquitous AI tools that perform standalone clinical tasks. The role of the physician likely remains paramount to clinical decision-making in the near future.

(JMIR Med Inform 2022;10(8):e34304) doi:10.2196/34304

KEYWORDS

medical education; artificial intelligence; health care trainees; AI; health care workers

Introduction

The field of artificial intelligence (AI) in medicine has seen rapid development in the last decade, with an increasing number of applications introduced in clinical settings [1]. With the rapid growth in computing power and data, medical AI has transformed from an afterthought into an imminent possibility. Currently, the utility of AI in completing tasks such as diagnostic prediction, automation, and generation of features from clinical data is recognized in many specialties. Models predicted the incidence of myocardial infarction and outperformed the current gold standard American College of Cardiology and American Heart Association risk algorithm [2]. These technological advancements have understandably raised concerns among health care trainees and professionals that AI may be taking over their duties. A study assessing medical students' views

regarding the impact of AI on future careers reported that 78.77% (1707/2167) expect significant changes due to AI and 89.62% (1942/2167) expressed that careful supervision by humans is required [3].

To moderate the concerns of AI in disrupting the future role of physicians, an understanding of the capabilities and limitations of AI tools is required. Wiens et al [4] reported AI adoption challenges, including problem formulation to market transition, all of which will require cooperation with interdisciplinary teams and systemwide change. In addition to refining the results of an AI algorithm, how the results are conveyed must also be accepted. Even if a physician accepts the judgement of a computer as legitimate, patients may not be nearly as receptive.

The aim of this commentary is to analyze the multifaceted issue of medical AI adoption to temper preconceived notions regarding its impact and rapid progression. We identify and explore four major barriers to AI adoption: (1) the limitations of performance and biases in AI applications, (2) the limitations due to heterogeneous digital infrastructure, (3) the limitations due to lack of technological literacy, and (4) the limitations of ethical challenges associated with medical AI usage.

Limitations of Performance

A significant barrier for AI applications to be implemented is regulatory approval, such as by the Food and Drug Administration (FDA), where AI applications would be included in the recently created category of Software as a Medical Device [5]. Certification is required for a recognized regulatory body to approve of a device's safety and effectiveness. If a new medical device is not considered a low- or moderate-risk device, it is required to enter the stringent premarket approval pathway, where demonstration of safety and effectiveness is required from clinical studies. The device is also classified in risk classes from Class I (the lowest risk) to Class III (the highest risk) [5]. AI, particularly machine learning, poses unique challenges as a machine learning model may continuously update with new training data. As such, the FDA has created recent guidelines, indicating that surveillance is required over the total product life cycle of the device, including model updates from retraining [6].

A standalone diagnostic tool would likely enter the premarket approval pathway and require extensive testing such as randomized controlled trials [7]. Leeuwen et al [8] evaluated 100 AI devices with CE-marked approval in Europe and reported that only 2 products were classified as class III, requiring premarket approval. Of 100 AI devices, 64 had no peer-reviewed studies validating the product performance. Wu et al [9] evaluated 54 AI medical devices approved by the FDA, with none being standalone diagnostic devices without physician supervision and none tested in a prospective trial. Hence, the current state of AI devices toward the FDA label of Computer-Assisted Detection Devices, which pose less resistance for market entry. The financial incentive results in a trend of devices being developed as physician-assisting tools that physicians can use at their discretion [10].

A technical barrier for AI devices to replace human analysis is the current performance of AI devices. For instance, when validated on a data set from a single center, convolutional neural networks (CNNs) routinely achieve accuracies above 0.90 [11]. However, with the variability of medical imaging from different machines, operators, or imaging protocols, multicenter studies are required to validate the generalizability of these classifiers. Alice et al [11] reported that 81% of diagnostic algorithms reviewed results in significant decrease of accuracy when externally validated. Thus, rigorous validation is required with a diverse data set to address the major machine learning challenges of data scarcity, population shifts from different data sets, prevalence shifts, and selection biases [12]. External validation also reveals a more accurate comparison between human and machine performance. Rodriguez-Ruiz et al [13] reported that when testing a published CNN to classify malignancies from mammography on a data set of 2652 images from seven different countries, the CNN performed within the same 95% CI accuracy range of 101 different radiologists [13].

The rigorous validation requirements for AI to be usable in clinical practice is evident when analyzing rapidly developed AI models. In the COVID-19 pandemic, over 100 diagnostic prediction models have been trained and published in literature, using features such as chest x-ray data, lung ultrasound, vital signs, and lab values. The reported concordance index of such models ranged from 0.71-0.99. However, Wynants et al [14] assessed that only 5% of the models found performed external validation, and only 2 models addressed selection biases during sampling.

An additional challenge for AI applications is that the ability to learn complex features is restricted to the architecture of the AI model. For instance, medical applications for CNNs commonly use architectures that perform well on the ImageNet challenge. The CNN architecture defines model parameters such as resolution, depth, and number of input channels, all of which affect the ability to detect complex features related to some objective. However, newer architectures are frequently developed, such as EfficientNet outperforming ResNet, DenseNet, Xception, and ResNeXT, all of which have been previously used in medical image classifiers [15]. Updating the model architecture is a significant change to the model. For instance, ResNet introduces the usage of residual blocks in a layer as an input for a subsequent layer to begin learning, changing how the model is initialized. This may require reapproval from regulatory bodies due to nontrivial changes in the device.

The alternative of a physician-assisting device is more likely in the near future, such as automating report extraction from imaging studies or image reconstruction to reduce excessive radiation from repeated imaging [16,17]. This reduces competition with physician tasks while still providing clinical utility from complex AI analyses.

Limitations of Current Infrastructure

Implementation of an AI product, even with validated performance, is limited by heterogenous digital infrastructure in health care systems. Different areas of patient care such as

inpatient progress notes, laboratory results, and discharge summaries may all have independent databases. This complexity is further multiplied by interactions with outpatient clinics and health authorities across provincial or state boundaries.

The incomplete adoption of electronic medical records (EMRs) illustrates the lag in digital infrastructure integration despite electronic record technology being available. The Canadian Federal Government's Economic Action Plan provided funding to health care providers toward establishing EMRs in primary care in 2010, leading to an increase of EMR adoption [18]. A similar progression took place in the United States in 2014 [19]. Despite this, there continues to be reliance on paper files in both primary care clinics and hospitals [20]. If, for instance, an algorithm in an emergency department requires baseline laboratory markers for a patient from their family physician, then standardization and likely digitalization of the input data is required.

There are currently 11 certified EMR vendors and 12 EMR products in Ontario [21]. Although hospitals often have a primary vendor, they often employ a variety of disparate EMR products in affiliated practices [21]. In theory, digitization of health care data would provide an abundance of high-quality data for AI research. However, EMR vendors operate in silos and use their own approach to storing data. To implement an AI product in practice may necessitate creation of a completely novel data pipeline to aggregate records across different databases. There are attempts at standardization including the "EMR Content Standard" by the Canadian Institute for Health Informatics [22]. This introduces a content standard for EMR data entry, but levels of prioritization of the standard differ across provinces, and no standard EMR data entry has been universally adopted, resulting in the persistence of difficulty in coalescing data to be usable by AI.

For AI technology to be successful, patients must consent to its use and trust the safety of the technology. A recent public opinion survey in the United States on AI indicated that data privacy was considered to be the most important issue [23]. Privacy concerns and restricted access limits access to a diverse and large sample size, which is necessary for an AI algorithm to be validated and implemented in clinical practice [24]. A diverse data set is also crucial to guarantee adequate representation of patient cohorts in AI algorithm training [25]. There are approaches to overcome these barriers including federated learning, where a model is shared across different centers for training without exporting data [24]. However, these approaches require universal agreements regarding scope and are currently not standard of practice.

Limitations of Technological Literacy

Medical AI applications have become increasingly relevant at an accelerated rate, though the lag in technological literacy of health care professionals for AI technology exceeds the expected social and cognitive lag of adapting new technology [26]. One challenge is that there is currently no standardized curriculum for AI education nor are there any relevant accreditation requirements within most medical doctorate programs [27]. This gap is significant as health care professionals are the main users of medical AI applications and will have to be responsible for appropriate usage of AI applications [28].

Despite a recent surge in interest in training health care trainees in AI, universal integration of AI education into current health care training is a nontrivial challenge. Medical training is dense and rigorous with significant demands on trainees and staff [29]. Implementation of such a curriculum also requires specific faculty expertise. Even with qualified educators available, there is the challenge of selecting the correct depth and breadth of topics required for medical trainees.

Without appropriate medical AI education, health care professionals may not be adequately equipped to navigate the potential ethical and legal implications of AI in health care. The flexibility that health care providers have in using their judgement to make clinical decisions tailored to an individual patient, using contextual understanding of interpatient and intrapatient variations, is essential to medicine. This process may be impeded if the end user lacks the basic digital literacy to understand the limitations of such applications of AI; for instance, deciding when to override an AI analysis in favor of contextual clinical judgement or vice versa. However, acquiring digital competency in AI applications may imply time away from service for health care providers and extra study workload for health care trainees, in addition to growing medical knowledge. Other challenges that contribute to the gap in technological literacy include lack of awareness of digital knowledge required for health care, lack of equitable access to AI education, and limited trust in AI applications in health care.

Medical applications must be well performing, trustworthy, transparent, interpretable, and explainable. Interpretation of AI models requires technical training, making it difficult to assess its performance. This is especially true in complex AI models such as deep neural networks, where it is not often possible to examine what features are used to compute the output, creating a colloquial "black box" algorithm. The gap in technological literacy among health care professionals, which is further hindered by the difficulty in implementing AI literacy training of an appropriate scope, prevents many AI applications from advancing beyond the proof-of-concept "computer-side" stage to bedside application [30].

Limitations of Ethical Challenges

In the presence of errors by AI decisions, there lies challenges not only in identifying liability but also in quality improvement analysis. Harm caused by AI may be due to several reasons in the pipeline, such as poor data stewardship, incomplete mathematical constraints resulting in an inaccurate model, or inappropriate usage by a clinician [31]. For instance, if an AI algorithm misdiagnoses a patient, causing an adverse event, is the error associated with data collection that was not representative of patient characteristics, with inadequate algorithm development resulting in computations that produce an inaccurate prediction, or with health care administration for deciding to use an AI product? Traditional quality improvement analysis in medicine, such as cause-effect analysis, may be insufficient because it lacks a 1-dimensional cause-to-effect pathway, particularly with multiparametric AI models such as

neural networks, which contain millions of computational kernels [32]. Interdisciplinary collaboration between data scientists, data stewards, clinicians, and health care workers is crucial to developing a risk liability and quality improvement system before AI can serve as a medical decision maker.

Additionally, substantial data bias may lead to unforeseen disparities in patient care as AI may stratify based on unintentional subgroups. Gichoya et al [33] observed that chest x-ray AI models can be used to predict patient's race with image features physicians were unaware of. The implication is that bias is unavoidable even when looking at data that appears agnostic, such as chest x-rays. This may further encourage health care disparities if the model makes decisions directly correlated with race or gender. There is then a utilitarian conflict of beneficence in deciding the extent to which it is acceptable to use an AI algorithm that may be more accurate and benefit certain subgroups at the expense of others; for instance, triaging resources for subgroups that AI can accurately analyze. There is also a deontological conflict to adhere to nonmaleficence. If we know there is a high likelihood of increasing disparity despite the beneficial aspects of AI, the application of AI would be unethical.

Hence, AI poses unique ethical issues due to limitations of transparency and inherent potential for harm when used as a decision maker. AI is capable of identifying hidden features within data that can be leveraged to improve decision-making, but it is not without potential risk and needs to be deliberated by all stakeholders involved in the process.

Conclusions

Implementation of AI in medicine faces barriers of regulatory approval, performance, compatibility of digital infrastructure, and shared multidisciplinary collaboration. Although AI shows potential in improving quality of life for patients by enhancing decision-making and tasks carried by health care professionals, the adoption of AI is likely incremental rather than a stark change in standard of care.

Conflicts of Interest

None declared.

References

1. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Front Med (Lausanne) 2020 Feb 5;7:27 [FREE Full text] [doi: 10.3389/fmed.2020.00027] [Medline: 32118012]
2. Deo RC. Machine learning in medicine. Circulation 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
3. Teng M, Singla R, Yau O, Lamoureux D, Gupta A, Hu Z, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. JMIR Med Educ 2022 Jan 31;8(1):e33390 [FREE Full text] [doi: 10.2196/33390] [Medline: 35099397]
4. 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 19;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
5. Smith JA, Abhari RE, Hussain Z, Heneghan C, Collins GS, Carr AJ. Industry ties and evidence in public comments on the FDA framework for modifications to artificial intelligence/machine learning-based medical devices: a cross sectional study. BMJ Open 2020 Oct 14;10(10):e039969 [FREE Full text] [doi: 10.1136/bmjopen-2020-039969] [Medline: 33055121]
6. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020 Sep 11;3(1):118 [FREE Full text] [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
7. Harvey HB, Gowda V. How the FDA regulates AI. Acad Radiol 2020 Jan;27(1):58-61. [doi: 10.1016/j.acra.2019.09.017] [Medline: 31818387]
8. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021 Jun 15;31(6):3797-3804 [FREE Full text] [doi: 10.1007/s00330-021-07892-z] [Medline: 33856519]
9. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021 Apr;27(4):582-584. [doi: 10.1038/s41591-021-01312-x] [Medline: 33820998]
10. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021 Apr 07;4(1):65 [FREE Full text] [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
11. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022 May 01;4(3):e210064 [FREE Full text] [doi: 10.1148/ryai.210064] [Medline: 35652114]
12. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun 2020 Jul 22;11(1):3673 [FREE Full text] [doi: 10.1038/s41467-020-17478-w] [Medline: 32699250]
13. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019 Mar 05:916-922. [doi: 10.1093/jnci/djy222] [Medline: 30834436]

14. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
15. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. Preprint posted online on May 28, 2019 . [doi: 10.48550/arXiv.1905.11946]
16. Carrodeguas E, Lacson R, Swanson W, Khorasani R. Use of machine learning to identify follow-up recommendations in radiology reports. J Am Coll Radiol 2019 Mar;16(3):336-343 [FREE Full text] [doi: 10.1016/j.jacr.2018.10.020] [Medline: 30600162]
17. Kambadakone A. Artificial intelligence and CT image reconstruction: potential of a new era in radiation dose reduction. J Am Coll Radiol 2020 May;17(5):649-651. [doi: 10.1016/j.jacr.2019.12.025] [Medline: 32004484]
18. Zhao EJ. The future of electronic medical records in Canada. CMAJ 2019 May 13;191(19):E542-E542 [FREE Full text] [doi: 10.1503/cmaj.71858] [Medline: 31085568]
19. Bristol N. The muddle of US electronic medical records. The Lancet 2005 May;365(9471):1610-1611. [doi: 10.1016/s0140-6736(05)66492-6]
20. Saleem JJ, Russ AL, Justice CF, Hagg H, Ebright PR, Woodbridge PA, et al. Exploring the persistence of paper with the electronic health record. Int J Med Inform 2009 Sep;78(9):618-628. [doi: 10.1016/j.ijmedinf.2009.04.001] [Medline: 19464231]
21. Larsen D, Hutchison S. Single electronic medical record for Canada: a second opinion. CMAJ 2019 May 13;191(19):E539-E540 [FREE Full text] [doi: 10.1503/cmaj.71810] [Medline: 31085566]
22. Keshavjee K, Williamson T, Martin K, Truant R, Aliarzadeh B, Ghany A, et al. Getting to usable EMR data. Can Fam Physician 2014 Apr;60(4):392 [FREE Full text] [Medline: 24733333]
23. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, AAO Task Force on Artificial Intelligence. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Transl Vis Sci Technol 2020 Aug 28;9(2):45 [FREE Full text] [doi: 10.1167/tvst.9.2.45] [Medline: 32879755]
24. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med 2020 Sep 14;3(1):119 [FREE Full text] [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]
25. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77 [FREE Full text] [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
26. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2:79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
27. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018 Sep 27;1(1):54 [FREE Full text] [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
28. 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]
29. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. J Intern Med 2018 Jun 24;283(6):516-529 [FREE Full text] [doi: 10.1111/joim.12752] [Medline: 29505159]
30. 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]
31. Smith H, Fotheringham K. Artificial intelligence in clinical decision-making: rethinking liability. Med Law Int 2020 Aug 26;20(2):131-154. [doi: 10.1177/0968533220945766]
32. Plsek PE. Quality improvement methods in clinical medicine. Pediatrics 1999 Jan;103(1 Suppl E):203-214. [Medline: 9917464]
33. Gichoya J, Banerjee I, Bhimireddy A, Burns J, Celi L, Chen L, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022 Jun;4(6):e406-e414 [FREE Full text] [doi: 10.1016/S2589-7500(22)00063-2] [Medline: 35568690]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
EMR: electronic medical record
FDA: Food and Drug Administration

Edited by C Lovis, J Hefner; submitted 16.10.21; peer-reviewed by A Joseph, E Ranschaert; comments to author 31.01.22; revised version received 29.07.22; accepted 02.08.22; published 15.08.22.

Please cite as:
Hu Z, Hu R, Yau O, Teng M, Wang P, Hu G, Singla R
Tempering Expectations on the Medical Artificial Intelligence Revolution: The Medical Trainee Viewpoint
JMIR Med Inform 2022;10(8):e34304
URL: https://medinform.jmir.org/2022/8/e34304
doi:10.2196/34304
PMID:35969464

Original Paper

Using the Diagnostic Odds Ratio to Select Patterns to Build an Interpretable Pattern-Based Classifier in a Clinical Domain: Multivariate Sequential Pattern Mining Study

Isidoro J Casanova[1], MSc; Manuel Campos[1,2,3], PhD; Jose M Juarez[1], PhD; Antonio Gomariz[4], PhD; Marta Lorente-Ros[5], MD; Jose A Lorente[6,7,8,9], MD, PhD

[1]AIKE Research Team (INTICO), Computer Science Faculty, University of Murcia, Murcia, Spain
[2]Murcian Bio-Health Institute (IMIB-Arrixaca), Murcia, Spain
[3]CIBERFES Fragilidad y Envejecimiento Saludable, Madrid, Spain
[4]Amazon Research, Madrid, Spain
[5]Department of Medicine, Mount Sinai St Luke's-Roosevelt Hospital, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[6]Intensive Care Unit, University Hospital of Getafe, Getafe, Spain
[7]School of Medicine, European University of Madrid, Madrid, Spain
[8]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain
[9]Department of Bioengineering, Universidad Carlos III, Madrid, Spain

Corresponding Author:
Isidoro J Casanova, MSc
AIKE Research Team (INTICO)
Computer Science Faculty
University of Murcia
Edificio 32, Campus de Espinardo
Murcia, 30100
Spain
Phone: 34 868887150
Email: isidoroj@um.es

Abstract

Background: It is important to exploit all available data on patients in settings such as intensive care burn units (ICBUs), where several variables are recorded over time. It is possible to take advantage of the multivariate patterns that model the evolution of patients to predict their survival. However, pattern discovery algorithms generate a large number of patterns, of which only some are relevant for classification.

Objective: We propose to use the diagnostic odds ratio (DOR) to select multivariate sequential patterns used in the classification in a clinical domain, rather than employing frequency properties.

Methods: We used data obtained from the ICBU at the University Hospital of Getafe, where 6 temporal variables for 465 patients were registered every day during 5 days, and to model the evolution of these clinical variables, we used multivariate sequential patterns by applying 2 different discretization methods for the continuous attributes. We compared 4 ways in which to employ the DOR for pattern selection: (1) we used it as a threshold to select patterns with a minimum DOR; (2) we selected patterns whose differential DORs are higher than a threshold with regard to their extensions; (3) we selected patterns whose DOR CIs do not overlap; and (4) we proposed the combination of threshold and nonoverlapping CIs to select the most discriminative patterns. As a baseline, we compared our proposals with Jumping Emerging Patterns, one of the most frequently used techniques for pattern selection that utilizes frequency properties.

Results: We have compared the number and length of the patterns eventually selected, classification performance, and pattern and model interpretability. We show that discretization has a great impact on the accuracy of the classification model, but that a trade-off must be found between classification accuracy and the physicians' capacity to interpret the patterns obtained. We have also identified that the experiments combining threshold and nonoverlapping CIs (Option 4) obtained the fewest number of patterns but also with the smallest size, thus implying the loss of an acceptable accuracy with regard to clinician interpretation. The best classification model according to the trade-off is a JRIP classifier with only 5 patterns (20 items) that was built using

unsupervised correlation preserving discretization and differential DOR in a beam search for the best pattern. It achieves a specificity of 56.32% and an area under the receiver operating characteristic curve of 0.767.

Conclusions: A method for the classification of patients' survival can benefit from the use of sequential patterns, as these patterns consider knowledge about the temporal evolution of the variables in the case of ICBU. We have proved that the DOR can be used in several ways, and that it is a suitable measure to select discriminative and interpretable quality patterns.

(JMIR Med Inform 2022;10(8):e32319) doi:10.2196/32319

KEYWORDS

sequential patterns; survival classification; diagnostic odds ratio; burn units

Introduction

Overview

Advances in the collection and storage of data have led to the emergence of complex temporal data sets, in which the data instances are traces of complex behavior characterized by time series of multiple variables.

In the clinical domain, patients who have incurred severe burns are treated in intensive care burn units (ICBUs). The first 5 days are fundamental: there is a resuscitation phase during the first 2 days and a stabilization phase during the following 3 days, and the patient's evolution (incomings, diuresis, fluid balance, pH, bicarbonate, base excess) is registered over this period. These variables are not considered in scores for mortality prediction and may play a relevant role in improving the current knowledge of the problem.

Designing algorithms that are capable of learning patterns and classification models from such data is one of the most challenging topics in data mining research [1]. One approach to deal with this problem is discovering patterns that are used as predictors in classification algorithms [2].

The number of patterns initially generated is usually very large, but only a few of these patterns are likely to be of interest to the domain expert that analyzes the data. There are several reasons for this: many of the patterns are either irrelevant or obvious, many patterns do not provide new knowledge regarding the domain, and many of them are similar or are included in others. Measures of the level of interest are, therefore, required to reduce the number of patterns, thus increasing the utility, usefulness, and relevance of the patterns discovered [3]. Some of these interestingness measures are based on the statistical significance of discriminative patterns.

In addition to traditional multidimensional analysis and data mining tasks, one interesting task is that of discovering notable changes and comparative differences. This leads to gradient mining and discriminant analysis [4].

Discriminative pattern mining is one of the most important techniques in data mining. This challenging task comprises a group of pattern mining techniques designed to discover a set of significant patterns that occur with disproportionate frequencies in different class-labeled data sets [5]. Research on discriminative patterns evolves rapidly under several nonuniform definitions, such as contrast sets, emerging patterns, or subgroups. However, these definitions are actually equivalent because their target patterns can be used interchangeably with the same ability to capture the differences between distinct classes [5].

The exploration of discriminative patterns generally includes 2 aspects: frequency and statistical significance. On the one hand, the frequency of a pattern can be assessed by its support, which is defined as the percentage of transactions (in our case, patients) that this pattern contains. A pattern is frequent if its support value is higher than a given threshold.

On the other hand, the statistical significance of discriminative patterns can be measured by using various statistic tests. A pattern is deemed significant if its significance value generated from a certain statistical measure could meet certain user-defined conditions, for example, no less (or more) than a given threshold. Any statistical measure that is capable of quantifying the differences between classes, such as the odds ratio, information gain, or chi-square, is generally applicable, and the choice of this measure will not typically affect the overall performance of the discriminative pattern discovery algorithms [5].

Many specific quantitative indicators of diagnostic test performance have been introduced into the clinical domain, such as sensitivity and specificity, positive and negative predictive values, chance-corrected measures of agreement, likelihood ratios or area under the receiver operating characteristic curve (AUC), among others. But there is a single indicator of diagnostic performance, denominated as the diagnostic odds ratio (DOR), which is closely linked to existing indicators, facilitates the formal meta-analysis of studies on diagnostic test performance, and is derived from logistic models [6].

We propose and compare 4 approaches in which the DOR is used as a statistical measure to select a reduced number of patterns, and we put forward the use of these patterns as predictors in a classification model. The calculation of the DOR for a pattern enables us to use a terminology that is closer to the language of clinicians, in which a pattern is considered to be a risk factor or to have a protection factor.

The first approach consists of using the DOR as a minimum threshold with which to select patterns. In the second approach, we calculate the difference in the DOR of a sequential pattern with respect to its extensions, and we establish a threshold for this difference to reduce the number of patterns selected. One advantage of this approach is that it can be used as an early pruning within the pattern discovery algorithm. In the third place, we calculate a CI for the DOR, and use this CI to prune patterns that are not statistically different from their extension

patterns. Finally, we combine the second and third approaches to select patterns with different properties.

We have verified that these propositions provide acceptable results by building a model for the classification of patients' survival using their daily evolution in an ICBU, employing multivariate sequential patterns. We have additionally compared the 4 approaches with the selection of patterns founded on classical frequency-based measures such as Jumping Emerging Patterns (JEPs).

Background

Sequential Pattern Mining

A sequence database is based on ordered elements or events, recorded with or without a concrete notion of time. There are many applications involving sequence data, such as economic and sales forecasting, speech or audio signals, web click streams, or biological sequences. The mining of frequently occurring ordered events or subsequences as patterns was first introduced by Agrawal and Srikant [7] and has become a significant challenge in data mining.

The purpose of sequential pattern mining is to discover interesting subsequences in a sequence database, that is, sequential relationships between items that are of interest to the user. Various measures can be used to estimate how interesting a subsequence is. In the original sequential pattern mining problem, the support measure is used. The support (or absolute support) of a sequence s in a sequence database is defined as the number of sequences that contain s, and is denoted by $sup(s)$.

Sequential pattern mining is the task of finding all the frequent subsequences in a sequence database. A sequence s is said to be a frequent sequence or a sequential pattern if and only if $sup(s) \geq minsup$, for a threshold $minsup$ established by the user. The assumption is that frequent subsequences are of interest to the user.

With regard to the algorithms employed to mine sequential patterns, there are 3 pioneer proposals: the GSP algorithm with the a priori strategy [8]; the SPADE algorithm, an a priori–based sequential pattern mining algorithm that uses vertical data format [9]; and PrefixSpan with the pattern growth strategy [10]. A number of algorithms based on these 3 proposals have focused on improving their efficiency using different search strategies or data structures.

The researchers refer the reader to [11] for more general information about sequential pattern mining.

Pattern and Sequence-Based Classification

Classification rule mining attempts to discover a small set of rules in the database to form an accurate classifier.

Initial approaches that combined pattern mining and classification models employed a strict stepwise approach, in which a set of patterns was computed once and those patterns were subsequently used in models. However, a large number of methods were later proposed, whose aim was to integrate pattern mining, feature selection, and model construction [12]. Some of these are Classification Based on Predictive Association Rules (CPAR), Classification Based on Multiple Association Rules (CMAR) [12], Multi-class, Multi-label Associative Classification (MMAC), and Classification Based on Associations (CBA). Many experimental studies have shown that these integrated classification methods have a high potential approach that builds more predictive and accurate classification systems than traditional classification methods such as decision trees [13].

The classification of sequence patterns is one of the most popular methodologies whose power has been demonstrated by multiple studies [14], and which has a broad range of real-world applications. In medical informatics, the classification of electrocardiogram time series (the time series of heart rates) shows whether the data originates from a healthy person or from a patient with heart disease [15], whereas in financial systems, transaction sequence data in a bank are classified for the purpose of fighting money laundering [16].

The sequence classification methods can be divided into 3 large categories [14]:

- The first category is that of feature-based classification, during which a sequence is transformed into a feature vector, after which conventional classification methods are applied. Feature selection plays an important role in this kind of methods.
- The second category is sequence distance–based classification. The distance function that measures the similarity between sequences determines the quality of the classification in a significant manner.
- The third category is model-based classification, such as using the hidden Markov model and other statistical models to classify sequences.

Conventional classification methods, such as neural networks or decision trees, are designed to classify feature vectors. One way to solve the problem of sequence classification is to transform a sequence into a vector of features by means of feature selections. Sequences can be classified by employing conventional classification methods, such as support vector machine and decision trees.

Several researchers have worked toward building sequence classifiers based on frequent sequential patterns. Lesh et al [17] proposed an algorithm for sequence classification using frequent patterns as features in the classifier. In their algorithm, subsequences are extracted and transformed into sets of features. After feature extraction, general classification algorithms such as support vector machine, naïve Bayes, or neural network can be used for classification. Their algorithm is the first attempt to combine classification and sequential pattern mining.

Tseng and Lee [18] proposed a Classify-By-Sequence (CBS) algorithm to combine sequential pattern mining and classification. Two algorithms, namely, "CBS Class" and "CBS All," were proposed in their paper. In "CBS Class," the database is divided into a number of subdatabases according to the class label of each instance. Sequential pattern mining is then implemented on each subdatabase. In "CBS All," a conventional

sequential pattern mining algorithm is applied on the whole data set. Weighted scoring is used in both algorithms.

With regard to the ICBU, few studies have dealt with the problem of survival prediction using machine learning or intelligent data analysis [19].

Interestingness Measures for Sequence Classification

In the original sequential pattern mining problem, the main measure used is support. The assumption is that frequent subsequences are of interest to the user.

A first important limitation of the traditional sequential pattern mining problem is that a huge number of patterns may be generated by the algorithms, depending on how the *minsup* threshold is set and on the characteristics of the database [11]. Finding too many patterns could hamper the effectiveness in some cases to which other measures could be better suited.

Many other rule interestingness measures are already used in data mining, machine learning, and statistics. Geng and Hamilton [20] have gathered together 9 different criteria that specify the interestingness of a pattern. These 9 criteria are conciseness, generality, reliability, peculiarity, diversity, novelty, surprisingness, utility, and actionability. These authors additionally classify these criteria into 3 main categories: objective, subjective, and semantics-based measures. Objective measures are those that depend only on raw data. Subjective measures are those that consider the users' background knowledge in addition to data, and finally semantic-based measures are a special type of subjective measures that take into account the explanation and the semantic of a pattern which are, like subjective measures, domain specific.

In this paper we focus on the probability-base objective measures used in the clinical domain. Some examples of objective rule interestingness measures that are often used in epidemiology as a statistical metric are presented in Table 1.

Table 1. Usual clinical objective rule interestingness measures for rules in the form of A→c.

Measure	Formula
Support	$P(Ac)$
Confidence	$P(c\|A)$
Coverage	$P(A)$
Prevalence	$P(B)$
Specificity	
Accuracy	
Diagnostic odds ratio	
Relative risk	

Relative risk and the DOR are statistical metrics that are often used in epidemiological studies. They are consistent: a larger odds ratio leads to a larger relative risk, and vice versa. Under the rare disease assumption, the DOR approximates the relative risk [21]. The DOR is usually used in case-control studies.

Li et al [21,22] used an epidemiological metric, relative risk, to measure pattern interestingness, and concluded that it is an optimal measure to find high-risk patterns. The proposed method was more efficient in covering the search space and produced a smaller number of rules. However, the number of rules in the output could still be too large for an easy interpretation. The authors applied the method to a real-world medical and pharmaceutical–linked data set and it revealed some patterns that are potentially useful in clinical practice.

Most of the conventional frequent pattern–based classification algorithms follow 2 steps [23]. The first step consists of mining a complete set of sequential patterns given a minimum support, while the second consists of selecting a number of discriminative patterns with which to build a classifier. In most cases, mining a complete set of sequential patterns in a large data set is extremely time-consuming, and the huge number of patterns discovered signifies that pattern selection and classifier building are also very time-consuming.

In fact, the most important consideration in sequence classification is not that of finding the complete rule set, but rather that of discovering the most discriminative patterns. In this respect, more attention has recently been paid to discriminative frequent pattern discovery for effective classification.

Heierman et al [24] presented a new data mining technique based on the Minimum Description Length principle, which discovers interesting features in a time-ordered sequence. Petitjean et al [25] introduced a method with which to exactly and efficiently identify the *k* most interesting patterns in a sequential database for which the difference between its observed and expected frequency is maximum: a measure denominated as leverage. Other authors focused on measures for the selection of patterns, such as the relative risk or a coverage measure [26].

In the clinical domain, univariate frequent episodes of Sequential Organ Failure Assessment (SOFA) subscores during the first days after admission were identified in Toma et al [27]. The

authors then selected a reduced number of patterns using Akaike's information criterion to build a logistic regression model to predict the survivability of patients with multiorgan failure. Later, Toma et al [28] showed that the use of univariate patterns as predictors is at least as effective as clinical scores.

After mining JEPs, Ghosh [29] used coupled hidden Markov learning models to build robust sequential patterns–based classifiers. This made it possible to predict hypotension risk, an acute hypotensive episode, or even of a septic shock, with the measurements of the mean arterial pressure, the heart rate, and the respiratory rate.

Survival Prediction in Intensive Care Burn Units

ICBUs are specialized units in which the main pathologies treated are inhalation injuries and severe burns. Early mortality prediction after admission is essential before an aggressive or conservative therapy can be recommended. Severity scores are simple but useful tools for physicians when evaluating the state of the patient [30]. Scoring systems aim to use the most predictive premorbid and injury factors to yield an expected likelihood of death for a given patient. Baux and Prognostic Burn Index scores provide a mortality rate by summing age and the percentage of total burn surface area, while the Abbreviated Burns Severity Index also considers gender and the presence of inhalation injuries.

The evolution of other parameters during the resuscitation phase (first 2 days) and during the stabilization phase (3 following days) may, however, also be important. The initial evaluation and resuscitation of patients with large burns that require inpatient care can be guided only loosely by formulas and rules. The inherent inaccuracy of formulas requires the continuous reevaluation and adjustment of infusions based on resuscitation targets. Incomings, diuresis, fluid balance, acid-base balance (pH, bicarbonate, base excess), and others help define objectives and assess the evolution and treatment response.

In the ICBU, a patient's evolution is registered but not considered in scores for mortality prediction. In a previous paper [31], we used emerging patterns with a knowledge-based temporal abstraction and then built classifiers of the survival of the patients with a high sensitivity and specificity. The results of the classification tests showed that our approach is comparable to the burn severity scores used currently by physicians.

Methods

Sequential Patterns

Let $I = \{i_1, i_2, ..., i_k\}$ be a set of items. An itemset is a non-empty subset of I. A sequence  is an ordered list of itemsets  (also called elements or events). Items within an element are unordered and would be listed alphabetically. An item can occur in an element of a sequence once at the most, but can occur multiple times in different elements of a sequence.

The number of instances of items in a sequence is denominated as the length of the sequence. A sequence with a length k is called a k-sequence. For example,  is a sequence that consists of 7 distinct items {a, b, c, d, e, f, g} and 6 itemsets. The length of the sequence is 12 items.

Each itemset in a sequence represents the set of events that occur at the same time (same timestamp). A different itemset appears at a different time.

Sequence  is a subsequence of sequence  (or β is a super-sequence of the sequence α), denoted as  , if there exist integers $i_1 < i_2 < \cdots < i_n$ such that  . For example,  is a subsequence of s.

The temporal representation of the patterns is principally carried out using time point representation or time interval representation.

In the time interval representation, there are different ways in which to relate intervals to each other, of which the best known is Allen's interval algebra [32] or the Time Series Knowledge Representation. In Allen's interval algebra, there are 13 relations that configure a very expressive language, thus making the pattern representation and the tasks related to temporal reasoning much more complicated.

Time point–based data are a special case of the time interval–based data, in which both the beginning and the end points occur at the same time (for each interval) and the relations between these points become simpler (before, equals or co-occurs, and after), usually denoted as (<, =, >). Furthermore, because the "after" operator (>) is the inverse of the "before" relation (<), if we always consider a relation from the point that occurs first, it is not necessary to use the "after" relation. For instance, if we have A>B, we will instead say B<A.

It is, therefore, possible to define patterns or sequences with only these 2 relations (<, =). Two patterns a and b are exactly equal if their points are exactly the same and they have exactly the same relations in the same positions, that is,  and  .

We have used the FaSPIP algorithm [33] to discover multivariate sequential patterns. FaSPIP is based on the equivalence classes strategy and is able to mine both points and intervals. Moreover, FaSPIP uses a new candidate generation algorithm based on boundary points and efficient methods to avoid the generation of useless candidates and to check their frequency.

In candidate generation, FaSPIP distinguishes between 2 operations to extend a sequence with an item, thus creating a new sequence: Sequence extensions (S-extensions), when the frequent points take place after, and Itemset extensions (I-extensions), when the points take place at the same time as the last item in the pattern. For instance, given the sequence  and a point  , the sequence  is an S-extension and  is an I-extension.

Emerging Patterns

The classical approach employed for pattern selection is based on the frequency of the patterns. Emerging patterns (EPs) or contrast sets are a type of knowledge pattern that describes significant changes (differences or trends) between 2 classes of

data [34]. EPs are sets of item conjunctions of attribute values whose frequency changes significantly from one data set to another. The problem of mining EPs can be expressed as follows: given 2 classes of data and a growth rate threshold, find all patterns (itemsets) whose growth rates—the ratio of their frequency between the 2 classes—are larger than the threshold [3].

Like other rules or patterns composed of conjunctive combinations of elements, EPs can be easily understood and used directly by clinicians.

Furthermore, the concept of JEPs [35] has been proposed to describe those discriminating features that occur only in the positive training instances but do not occur in the negative class at all. The most frequently appearing JEPs have been used to build accurate classifiers [36,37].

Diagnostic Odds Ratio and CI

Clinicians must rely on the correct interpretation of diagnostic data in a variety of clinical environments. A 2×2 table is an essential tool to present the data regarding epidemiological studies for diagnostic test evaluation (Table 2). The terms commonly used with diagnostic tests are sensitivity, specificity, and accuracy, which statistically measure the performance of the test. *Sensitivity* indicates how well the test predicts one category and *specificity* measures how well the test predicts the other category, while *accuracy* is expected to measure how well the test predicts both categories.

Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)

Other multiple tests with which to improve diagnostic decision making in different clinical situations have also been suggested. For example, Glas et al [6] proposed the use of the DOR as a single indicator of diagnostic performance.

Table 2. 2×2 Contingency table.

Test	Reference test	
	Target disorder	No target disorder
Positive	TP[a]	FP[b]
Negative	FN[c]	TN[d]

[a]TP: true positive.

[b]FP: false positive.

[c]FN: false negative.

[d]TN: true negative.

The DOR is used to measure the discriminative power of a diagnostic test: the ratio of the odds of a positive test result among the diseased to the odds of a positive test result among the nondiseased. The DOR is not prevalence dependent, and may be easier to understand, as it is a familiar epidemiological measure. It can be expressed in terms of sensitivity and specificity.

DOR = (TP/FN)/(FP/TN) = [sensitivity / (1–sensitivity)] / [(1–specificity) / specificity]

The value of a DOR ranges from 0 to infinity. To calculate the DOR, the potential problems involving division by 0 are solved by adding 0.5 to the selected cells in the diagnostic 2×2 table.

The further the odds ratio is from 1, the more likely it is that those with the disease are exposed when compared with those without the disease (risk factor). A value of 1 means that a test does not discriminate between patients with the disorder and those without it. Values lower than 1 suggest a reduced risk of disease associated with exposure (protection factor).

CIs for range estimates can be conventionally calculated as shown in the next equation:

where *Xhm* is the Mantel-Haenszel chi-square and *Z*=1.96 if a confidence of 95% is employed.

Li et al [38] built an algorithm based on the following assumption: if adding an exposure to a rule does not produce a significant change in the DOR, then the rule should not be reported. The DOR between 2 rules is significantly different if their 95% CIs do not overlap.

Several studies based on the nonoverlapping of the DOR have been performed. Toti et al [39] discussed the differences in performance achieved while extracting rules with the different definitions of a nonexposed population, when no pruning criterion is used to filter redundant rules, or when a pruning criterion of redundant rules based on overlapping of 95% CI is added. They confirmed that mining with no pruning criterion produces a high number of redundant rules, thus proving the need for a process with which to eliminate them. Toti et al [40] in another study explained that the traditional interest metrics of support and confidence need to be substituted for metrics that focus on risk variations caused by different exposures. They proposed 2 postprocessing pruning criteria: a rule is pruned if its 95% CI for the DOR crosses the value of 1 or if there is no overlapping of the 95% CI of the rule with all of its parents.

Case Study

A database contains 480 patient registries, which were recorded between 1992 and 2002. In this database, the temporal attributes that allow the monitoring and evaluation of the response to the treatment of patients are recorded once a day for 5 days. All attributes are continuous variables and represent the value

accumulated during 24 hours. The registered variables are (1) total of managed liquids measured in cubic centimeters (cc) represented in the patterns as *INC*; (2) diuresis in cubic decimeters (dc) represented in the patterns as *DIUR*; (3) balance of fluids in cubic decimeters (dc) represented in the patterns as *BAL*; (4) pH; (5) bicarbonate in millimoles/liter (mmol/L) represented in the patterns as *BIC*; and (6) excess base in milliequivalents/liter (mEq/L) represented in the patterns as *BE*. Note that fluid balance is not the difference between revenues and diuresis, but is rather considered to be all the possible eliminations of fluids.

We have removed from the database only those patients who died during the course of the study or those for whom it was not possible to estimate the duration of their hospital stay. After this cleansing, 465 patients remained, of whom 378 patients (81.3%) survived, 324 patients (69.7%) were male, and 201 patients (43.2%) had inhalation injuries. Table 3 provides a summary of the static attributes of the database.

Table 3. Attribute summary.

Attribute	Minimum	Maximum	Median	SD
Age (years)	9	95	46.42	20.34
Weight (kg)	25	120	71.05	10.77
Length of stay (days)	3	162	25.02	24.24
Total burn surface area (%)	1	90	31.28	20.16
Deep burn surface area (%)	0	90	17.01	17.41
Simplified Acute Physiology Score	6	58	20.67	9.49

Experiments

We carried out the experiments by following the 4-step knowledge discovery process described in our previous paper [31]: (1) preprocessing, (2) mining, (3) pattern selection, and (4) classification.

In the first step, the preprocessing was carried out by employing 2 different discretization methods for the continuous attributes. One method was attribute discretization performed by an expert. This method provided the patterns with greater interpretability, because they are expressed in clinical language. The other method is the unsupervised correlation preserving discretization (UCPD), because it provided the best classification in comparison to several automatic discretization algorithms [41].

In the second step, we used the FaSPIP algorithm [33] to discover multivariate sequential patterns. We considered pattern supports ranging from 16% to 6% to find the greatest support that generates the smallest number of patterns with the best classification results. This, therefore, enabled us to obtain interesting patterns, ranging from a small number to thousands of them (Table 4).

The best results were not produced with the lowest supports, which seems to imply that there is no overfitting.

The third step consisted of reducing the number of patterns found to select only those that would be relevant for the classification. If the support used in the previous step is low, the number of frequent patterns increases acutely: the pattern explosion phenomenon is one important disadvantage of using patterns as predictors for classifiers.

We decided to use a baseline experiment to compare it with our proposed methods. We therefore employed the frequency property (because it is frequently used to measure interestingness) to select discriminative patterns. To this end, we selected only JEPs that are not common in the subset of nonsurvivors and survivors, thus enabling us to remove common behavior or a patient's evolution that is not discriminative.

Finally, the fourth step consisted of building a classification model with the constraint that it had to be interpretable. We wished to obtain a model with a small number of patterns that would be easy for the physician to interpret. In this case, we used a rule learner and a decision tree.

On the one hand, we used Repeated Incremental Pruning to Produce Error Reduction (RIPPER) as a rule learner. With this sequential covering algorithm, rules are learned one at a time, and each time a rule is learned, the tuples covered by the rule are removed. This process is repeated until there are no more training examples or if the quality of a rule obtained is below a user-specified threshold. JRIP (the implementation of RIPPER in WEKA) is one of the best classification algorithms to combine human readability and accuracy [42].

On the other hand, we choose the J48 decision tree implemented by WEKA for the C4.5 algorithm. This employs a greedy technique that is a variant of ID3, which determines the most predictive attribute in each step, and splits a node based on this attribute. Mohamed et al [43] explained that J48 produces high accuracy of classification and simple tree structure. Moreover, Jiménez et al [19] showed that the J48 decision tree algorithm provides the simplest model using the ICBU data set, and thus it is easily interpretable by physicians.

In all cases, we configured the classifiers with the same minimum number of elements in each leaf to 2% and also with the minimal weights of rule instances within a split to 2%. The accuracy, sensitivity, specificity, and AUC were calculated using a 10-fold cross validation.

Table 4. Number of interesting patterns selected after mining on the subset of survivors and on the set of nonsurvivors for UCPD[a] and expert discretization

Discretization and support (%)	Survival + death initial patterns	Baseline JEPs[b]	Experiment 1, DOR[c]		Experiment 2, differential DOR		Experiment 3, nonoverlapping DOR		Experiment 4, differential + nonoverlapping DOR	
			<.08, >16	<.04, >32	All	Best	All	Best	All	Best
Expert										
10	46,041 + 83,015	391	2065	750	2795	2359	858	746	236	198
8	88,084 + 241,866	4931	14,424	5798	10,655	8781	2195	1856	701	504
6	224,952 + 492,504	47,113	51,352	41,059	32,406	26,157	4545	3803	1556	1293
UCPD										
16	238,337 + 49,947	2179	14,158	2766	2401	1990	1529	1415	325	272
14	396,238 + 68,654	7556	33,979	7483	4153	3465	2296	2052	487	411
12	647,943 + 137,546	22,940	65,564	16,272	9907	8173	6418	5228	1397	1212

[a]UCPD: unsupervised correlation preserving discretization.

[b]JEP: Jumping Emerging Pattern.

[c]DOR: diagnostic odds ratio.

Ethics Approval

The study was approved by the Ethics Committee of Hospital Universitario de Getafe (38/17, approved on 30/11/2017). This research study was conducted from data obtained for clinical purposes. Informed consent was not required.

Results

Overview

The results of the baseline experiment and the results of our 4 different proposals using the DOR are shown below. The number of patterns generated in the subset of survivors and in the set of nonsurvivors with different supports is shown in Table 4. We also studied the length of the patterns produced (Table 5). A short pattern is simpler and more general (it covers more patients). However, a long pattern is more specific (covers fewer patients) and is harder to understand. It is, therefore, more difficult to build a classifier with short patterns.

In the discussion, we explore 3 aspects: classification performance, number and length of patterns selected, and classification interpretability.

Table 5. Number (and percentage) of interesting patterns by length (from 2 to 10) for 8% expert discretization and selecting all the patterns when it is possible.

Pattern length	Baseline JEPs[a] (n=4931)	Experiment 1a, DOR[b] (<0.08, >16) (n=14,424)	Experiment 1b, DOR (<0.04, >32) (n=5798)	Experiment 2, differential DOR (n=10,655)	Experiment 3, nonoverlapping DOR (n=2195)	Experiment 4, differential + nonoverlapping DOR (n=701)
2	0 (0)	5 (0.0)	0 (0)	289 (2.7)	76 (3.5)	39 (5.6)
3	41 (0.8)	187 (1.3)	49 (0.8)	2063 (19.4)	461 (21.0)	198 (28.2)
4	542 (11.0)	1610 (11.2)	552 (9.5)	3912 (36.7)	857 (39.0)	299 (42.7)
5	1377 (27.9)	4176 (29.0)	1545 (26.6)	3004 (28.2)	612 (27.9)	140 (20.0)
6	1518 (30.8)	4811 (33.4)	1960 (33.8)	1155 (10.8)	175 (8.0)	23 (3.3)
7	987 (20.0)	2698 (18.7)	1190 (20.5)	212 (2)	14 (0.6)	2 (0.3)
8	372 (7.5)	785 (5.4)	407 (7.0)	20 (0.2)	0 (0)	0 (0)
9	84 (1.7)	139 (1.0)	85 (1.5)	0 (0)	0 (0)	0 (0)
10	10 (0.2)	13 (0.1)	10 (0.2)	0 (0)	0 (0)	0 (0)

[a]JEP: Jumping Emerging Pattern.

[b]DOR: diagnostic odds ratio.

Baseline Experiment: Using JEPs

In the baseline experiment, we searched for discriminative patterns, one of the most important techniques in data mining [44], where the patterns are pruned using only support properties. We selected JEPs, signifying that we maintained patterns found only in the survivors and patterns that occurred exclusively in the nonsurvivors. In a previous paper [31], we verified that this type of emerging patterns produces the best

classification results. Furthermore, in this way there is no need to set a threshold that could bring out different results.

Table 6 depicts the results of the experiments carried out using 2 discretization algorithms and by varying the pattern support.

Table 6. Results of the baseline experiment with JEPs.[a,b]

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[c]
J48							
Expert							
10	7	33	4.71	100.00	43.68	89.46	0.709
8	*17*	*84*	*4.94*	*100.00*	*56.32*	*91.83*	*0.782*
6	16	80	5	100.00	44.83	89.68	0.720
UCPD[d]							
16	8	29	3.63	100.00	52.87	91.18	0.763
14	*10*	*37*	*3.7*	*100.00*	*66.67*	*93.76*	*0.853*
12	12	48	4	100.00	59.77	92.47	0.796
JRIP							
Expert							
10	8	37	4.63	100.00	40.23	88.82	0.704
8	*15*	*79*	*5.27*	*100.00*	*58.62*	*92.26*	*0.777*
6	18	87	4.83	100.00	44.83	89.68	0.729
UCPD							
16	7	34	4.86	100.00	47.13	90.11	0.711
14	*10*	*35*	*3.5*	*100.00*	*73.56*	*95.05*	*0.866*
12	12	51	4.25	100.00	62.07	92.90	0.833

[a]JEP: Jumping Emerging Pattern.

[b]Highest specificity is in italics.

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

[d]UCPD: unsupervised correlation preserving discretization.

As will be noted, the JEPs make it possible to achieve a sensitivity of 100%, but the specificity has lower values. This is due to the fact that the data set is imbalanced with a majority of survivors, and the patterns cover only those patients that will survive or those that will die. It is necessary to achieve a higher specificity to predict the nonsurvivors, so the highest specificity is in italics in Table 6 as a baseline best result.

The expert discretization is preferred by clinicians, because it is based principally on reference ranges values. But note that it is possible to improve the results by using an automatic discretization, such as UCPD (see [41]).

When using expert discretization, the highest specificity (58.62%) is obtained using the JRIP classifier with 8% support.

This classifier requires 15 patterns, with a total length of 79 items, with the average length per pattern being 5.27 items. As an example, we show a pattern found in the subset of nonsurvivors. For each variable, the subindex i marks the i discretization interval where i=0 is the lowest interval:

$< BAL_4 < BIC_1 < DIUR_2 < BE_0$ (10 nonsurvivors, 0 survivors)

There is also an interesting pattern that appears in all the 5 experiments for the subset of nonsurvivors:

$< DIUR_3 < INC_0 < INC_0 < DIUR_3$ (10 nonsurvivors, 0 survivors)

It would, therefore, be possible to interpret this pattern as “a patient will die if his/her diuresis is very high on one day, and during the next 2 days there is a low income with a very high diuresis the following day.”

Experiment 1: Using the DOR

In this experiment, we calculated the DOR for each pattern as shown in “Methods” section. In clinical language, a DOR>1 implies that the exposure to the pattern is a risk factor. Conversely, a DOR<1 implies that the pattern is a protection factor and selecting a DOR threshold with a very low value therefore suggests a reduced risk of disease associated with exposure. A value of DOR=1 signifies that the pattern does not discriminate between patients with the disorder and those without it.

The selection of patterns with either a high value or a low value for the DOR will therefore generate discriminative patterns. It

is necessary to establish a manual threshold for the value of the DOR to choose the patterns. We have carried out 2 experiments. In the first experiment (1a), we have selected the patterns with a DOR value higher than 16 or lower than 0.08, and in the second experiment (1b), we have selected more exigent values, which were double or half the DOR value, that is, with a DOR value higher than 32 or lower than 0.04. This allowed us to reduce the number of patterns (Table 4) and we obtained a number of patterns in Experiment 1b that were similar to those obtained in the previous experiment. In the more exigent configuration, the length of the selected patterns was almost 6 (Table 5), which was again similar to the baseline experiment.

Tables 7 and 8 show the classification performance of the 2 experiments using expert discretization and UCPD methods with different pattern supports. Expert discretization makes it possible to attain better results than when using JEPs in the previous experiment (Table 6), and worse results than when using UCPD.

Table 7. Results of Experiment 1a using the DOR[a] (<0.08, >16).

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[b]
J48							
Expert							
10	13	67	5.15	90.21	62.07	84.95	0.766
8	18	89	4.94	88.62	58.62	83.01	0.759
6	16	80	5	91.80	47.13	83.44	0.702
UCPD[c]							
16	8	29	3.62	100.00	52.87	91.18	0.763
14	11	43	3.91	100.00	62.07	92.90	0.787
12	12	48	4	100.00	59.77	92.47	0.796
JRIP							
Expert							
10	10	46	4.6	91.27	55.17	84.52	0.716
8	12	58	4.83	93.12	54.02	85.81	0.720
6	14	67	4.79	94.44	52.87	86.67	0.706
UCPD							
16	8	33	4.13	100.00	41.38	89.03	0.716
14	12	47	3.92	100.00	62.07	92.90	0.828
12	12	46	3.83	100.00	59.77	92.47	0.816

[a]DOR: diagnostic odds ratio.

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

[c]UCPD: unsupervised correlation preserving discretization.

Table 8. Results of Experiment 1b using the DOR[a] (<0.04, >32).

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[b]
J48							
Expert							
10	10	49	4.9	93.65	50.57	85.59	0.710
8	17	84	4.94	94.18	55.17	86.88	0.767
6	16	80	5	95.50	37.93	84.73	0.656
UCPD[c]							
16	8	29	3.62	100.00	52.87	91.18	0.763
14	11	43	3.91	100.00	62.07	92.90	0.787
12	12	48	4	100.00	59.77	92.47	0.796
JRIP							
Expert							
10	11	50	4.55	97.09	44.83	87.31	0.704
8	14	67	4.79	95.50	62.07	89.25	0.801
6	16	87	5.44	98.15	48.28	88.82	0.715
UCPD							
16	7	26	3.71	100.00	47.13	90.11	0.727
14	11	45	4.09	100.00	60.92	92.69	0.792
12	14	55	3.93	100.00	60.92	92.69	0.822

[a]DOR: diagnostic odds ratio.

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

[c]UCPD: unsupervised correlation preserving discretization.

If we choose expert discretization, with a JRIP classifier and the highest values of the DOR (Table 8), we obtain a higher specificity than with JEPs (62.07%), but a lower sensitivity (95.50%). This can be explained as follows: if we look at one of the 14 patterns used in that classifier, we can find an example of a short pattern with only 3 items:

$BIC_1 < BAL_4 < PH_1$ (72.30 DOR) (14 nonsurvivors, 1 survivor)

This pattern, with a DOR value of 72.30, classifies a group of patients that will die, although we know that there will be minimal errors (1 patient survives).

We selected the pattern $DIUR_3 < INC_0 < INC_0 < DIUR_3$ in this experiment because it has a DOR value of 98.05, and it is necessary to recall that all the patients in this pattern will die (10 deaths, 0 survivors). This kind of JEP therefore produces a good specificity, and consequently 100% sensitivity (there are no classification errors).

Experiment 2: Using the Differential DOR Between a Pattern and Its Extensions

A sequential pattern p_i, of a specific length (l), in a point in time (t), has a DOR value $DOR(p_i)$. In every extension of this pattern (l+1), which could be an S-extension (in the next time, t+1) or an I-extension (in the same time, t), there will be n several patterns (p_{i1}, p_{i2}, ..., p_{in}) that are children of super-pattern p_i with different DOR values, . In this experiment, we choose only the patterns that had a difference in DOR value between the super-pattern and its extensions higher than a threshold γ, that is $DOR(p_i) - DOR(p_{ij}) > \gamma$.

For a better interpretation of the DOR, we calculated the risk factor probability $R(p_i)$ and the protection factor probability $P(p_i)$ as shown in the next equations:

$$R(p_i) = DOR(p_i)/[DOR(p_i) + 1]$$

$$P(p_i) = 1 - R(p_i)$$

In our experiment we, therefore, selected the patterns with 2 conditions: (1) when the difference between the risk factor probability $R(p_i)$ was greater than 25% or (2) when the difference between the protection factor probability $P(p_i)$ was greater than 30%. We chose a lower threshold value for $R(p_i)$ because we wished to obtain a higher specificity by having more patterns that were representative of nonsurvivors. In this experiment we obtained patterns with a high quality that produced great changes in the evolution of the patients.

We additionally used 2 alternative strategies to select patterns: it is possible to maintain all the extensions with a difference in the DOR value that is higher than a threshold or to explore the extensions with a beam search, in which case we select only the most promising extension with the highest DOR difference

among all extensions. Tables 9 and 10 show the results attained using both strategies.

With regard to the number of patterns selected (Table 4), when we have chosen the best extension, we have only reduced the total number of patterns by less than one-third because the majority of the patterns only have 1 or 2 extensions.

If we study the length of the patterns (Table 5), in this experiment (and in those that follow) the majority of the patterns have a length of around 4, and it is now possible to find more patterns with a shorter length. Note that the distribution of patterns by length has changed. We currently have more general patterns that are shorter. This produces worse classification results when we use expert discretization with a JRIP classifier. It is well known that expert discretization usually performs worse because it is not based on a statistical or information theory that has been specifically designed for classification purposes. This also occurs in almost all of the following experiments.

However, the results obtained with UCPD are similar, and even with the JRIP classification and beam search, we need the lowest number of items and patterns from all the experiments: only 5 patterns with a total length of 20 items are required to attain 56.32% specificity.

Table 9. Results of Experiment 2a using the differential DOR[a] (keeping all pattern extensions).

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[b]
J48							
Expert							
10	28	100	3.57	89.42	49.43	81.94	0.662
8	21	89	4.24	86.51	62.07	81.94	0.773
6	18	84	4.67	96.30	44.83	86.67	0.694
UCPD[c]							
16	21	81	3.86	93.65	49.43	85.38	0.677
14	15	56	3.73	94.97	56.32	87.74	0.759
12	12	52	4.33	100.00	58.62	92.26	0.788
JRIP							
Expert							
10	4	13	3.25	90.74	31.03	79.57	0.620
8	8	25	3.13	86.77	29.89	76.13	0.600
6	3	7	2.33	89.68	29.89	78.49	0.594
UCPD							
16	10	37	3.70	92.86	24.14	80.00	0.594
14	11	41	3.73	94.18	33.33	82.80	0.674
12	8	26	3.25	96.03	62.07	89.68	0.831

[a]DOR: diagnostic odds ratio.

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

[c]UCPD: unsupervised correlation preserving discretization.

Table 10. Results of Experiment 2b using the differential DOR[a] (using beam search for best pattern extension).

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[b]
J48							
Expert							
10	20	73	3.65	89.15	44.83	80.86	0.642
8	21	88	4.19	87.57	62.07	82.80	0.783
6	18	84	4.67	97.35	43.68	87.31	0.710
UCPD[c]							
16	21	81	3.86	93.65	49.43	85.38	0.675
14	15	56	3.73	94.71	56.32	87.53	0.760
12	12	52	4.33	100.00	57.47	92.04	0.764
JRIP							
Expert							
10	18	59	3.28	89.15	27.59	77.63	0.582
8	5	17	3.4	90.48	21.84	77.63	0.569
6	8	29	3.62	91.53	31.03	80.22	0.623
UCPD							
16	9	31	3.44	91.01	28.74	79.35	0.618
14	19	71	3.74	94.18	34.48	83.01	0.683
12	5	20	4	97.09	56.32	89.46	0.767

[a]DOR: diagnostic odds ratio.

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

[c]UCPD: unsupervised correlation preserving discretization.

The J48 classification tree used to classify with expert discretization and 8% support, using beam search for the best pattern extension, makes it possible to attain 62.07% specificity, and require 21 patterns, with an average length of 4.19 items per pattern. This average is the lowest value of all the experiments carried out using the J48 classifier with expert discretization. Within these 21 patterns, we can find 2 patterns with only 2 items, which are used to classify the survivors:

$DIUR_3 < BE_2$ (40.23% PROTECTION) (43 deaths, 150 survivors)

$INC_2 = PH_3$ (43.58% PROTECTION) (35 deaths, 176 survivors)

The first pattern, $DIUR_3 < BE_2$, is interesting because if the *PH* is very high the next day and has the extension $DIUR_3 < BE_2 < PH_4$ (78.85% PROTECTION; 5 deaths, 70 survivors), the patient survival rate increases by 38.62%.

Furthermore, we have discovered a pattern with which to classify the nonsurvivors that can also be found in the J48 tree classifiers of the subsequent experiments, and that was not selected in the classification algorithms used in the previous experiments:

$p_{i1} = BIC_1 < BIC_1 < PH_1$ (98.87% RISK; 9 deaths, 0 survivors)

This pattern has a DOR value of DOR(p_{i1}) = 87.12, with a risk probability of $R(p_{i1})$ = 98.87%. It has been selected because its super-pattern $p_i = BIC_1 < BIC_2$ (44 deaths, 111 survivors) has a DOR value of DOR(p_{i1}) = 2.46, with a risk probability of $R(p_i)$ = 71.1%. This signifies that there is an increase in the risk of $R(p_{i1}) - R(p_{i1})$ = 27.77%, which is higher than the 25% fixed threshold.

Experiment 3: Using the Nonoverlapping of the CI of the DOR

In this experiment, we have selected patterns based on the nonoverlapping of 95% CI of the DOR (as stated in [38]). In addition, only patterns whose CI does not include the value 1 have been included in the output (as occurred in [40]). All the patterns are, therefore, either a protector factor or a risk factor, but not both or undetermined.

Table 11 shows the results obtained when we maintain all the pattern extensions, while Table 12 shows the results obtained when only the best pattern extension is chosen using beam search.

We also obtain a reduced number of patterns with respect to the previous experiment (Table 4), and an advantage of this experiment is that this number does not depend on a threshold value.

In general, the classification performance is similar to that of the previous experiments, although with the JRIP classification using expert discretization, we obtain better results when selecting only the best child.

The J48 classification tree used to classify with expert discretization, and 8% support, using beam search for best pattern extension, allows us to obtain 58.62% specificity and a higher sensitivity than the previous experiment: 16 patterns are required.

One of the shortest patterns that we find in the J48 classification tree is:

$PH_4 < PH_4 < BE_1$ (6 deaths, 1 survivors)

This pattern has a DOR value of 27.93 in the interval (6.71, 116.26). Its super-pattern $PH_4 < PH_4$ (14 deaths, 109 survivors) has a DOR value of 0.47 in the interval (0.26, 0.87). Note that the CI of these patterns does not overlap.

Table 11. Results of Experiment 3a using the nonoverlapping CI of DOR[a] (keeping all pattern extensions).

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[b]
J48							
Expert							
10	10	41	4.1	93.92	48.28	85.38	0.721
8	16	77	4.81	94.97	58.62	88.17	0.741
6	18	90	5	96.56	56.32	89.03	0.768
UCPD[c]							
16	18	70	3.89	97.35	57.47	89.89	0.794
14	11	43	3.91	99.74	62.07	92.69	0.803
12	11	47	4.27	100.00	57.47	92.04	0.786
JRIP							
Expert							
10	11	37	3.36	93.65	41.38	83.87	0.682
8	13	60	4.62	91.80	33.33	80.86	0.641
6	7	30	4.29	96.56	42.53	86.45	0.722
UCPD							
16	6	23	3.83	96.30	41.38	86.02	0.727
14	9	33	3.67	98.94	56.32	90.97	0.803
12	14	58	4.14	96.30	60.92	89.68	0.793

[a]DOR: diagnostic odds ratio.

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

[c]UCPD: unsupervised correlation preserving discretization.

Table 12. Results of Experiment 3b using the nonoverlapping CI of DOR[a] (using beam search for best pattern extension).

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[b]
J48							
Expert							
10	10	41	4.1	94.18	51.72	86.24	0.742
8	16	77	4.81	94.71	58.62	87.96	0.739
6	18	90	5	96.83	55.17	89.03	0.758
UCPD[c]							
16	16	68	4.25	96.30	55.17	88.60	0.798
14	13	51	3.92	100.00	62.07	92.90	0.795
12	11	45	4.09	100.00	60.92	92.69	0.812
JRIP							
Expert							
10	6	20	3.33	94.44	48.28	85.81	0.735
8	16	62	3.88	95.24	41.38	85.16	0.700
6	12	51	4.25	95.77	52.87	87.74	0.747
UCPD							
16	16	66	4.13	95.50	40.23	85.16	0.695
14	12	44	3.67	97.88	54.02	89.68	0.747
12	15	60	4	99.21	55.17	90.97	0.788

[a]DOR: diagnostic odds ratio.

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

[c]UCPD: unsupervised correlation preserving discretization.

Experiment 4: Using a Differential DOR With the Nonoverlapping of the CI

The last proposal consists of using the previous 2 approaches together (Experiments 2 and 3), signifying that we prune the patterns based on the overlapping of the CI of the DOR, and also based on the difference between the risk (or protection) factor probabilities. In both cases, we maintain the same thresholds.

In this experiment we substantially reduced the number of patterns generated (Table 4). For example, in the case of expert discretization and 8% support (keeping all pattern extensions), we obtained only 701 patterns with this experiment, which is a decrease of 68.06% from nonoverlapping DOR (with 2195 patterns) and a decrease of 85.78% with respect to the baseline experiment (with 4931 patterns).

It is necessary to consider that if the number of patterns is too low, we do not usually achieve a good classification result. But with this experiment, for example, with 8% support, expert discretization, and the J48 classifier, with only 504 patterns, we have obtained a similar result to previous ones, using only 13 patterns in the classifier, with a sensitivity of 96.30% and a specificity of 57.47% in the beam search for the best pattern extension (Table 13). This is the lowest number of patterns required for expert and J48 discretization, with a total length of only 55 items.

The classification performance, as is shown in Tables 13 and 14, is similar to that of the previous experiments.

Let us now analyze the pattern that is selected in this experiment and in all the previous experiments: $DIUR_3 < INC_0 < INC_0 < DIUR_3$ (10 deaths, 0 survivors). It has a DOR value of 98.05 in the interval (24.21, 397.18), with a risk probability of 98.99%. Its super-pattern $DIUR_3 < INC_0 < INC_0$ has a DOR value of 2.07 in the interval (1.20, 3.57) with a risk probability of 67.39%, signifying that there is no overlapping in the CI, and that there is an increase in the risk probability of 31.6%.

Table 13. Results of Experiment 4b using the differential DOR[a] and the nonoverlapping CI (using beam search for best pattern extension).

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[b]
J48							
Expert							
10	10	35	3.5	95.50	41.38	85.38	0.694
8	13	55	4.23	96.30	57.47	89.03	0.770
6	16	75	4.69	98.41	50.57	89.46	0.739
UCPD[c]							
16	20	74	3.7	93.92	50.57	85.81	0.758
14	7	28	4	96.83	58.62	89.68	0.808
12	12	50	4.17	100.00	59.77	92.47	0.812
JRIP							
Expert							
10	6	21	3.5	92.59	25.29	80.00	0.597
8	14	43	3.07	91.80	29.89	80.22	0.614
6	15	57	3.8	92.59	29.89	80.86	0.626
UCPD							
16	10	37	3.7	96.83	35.63	85.38	0.671
14	10	36	3.6	98.68	32.18	86.24	0.673
12	15	59	3.93	98.68	50.57	89.68	0.759

[a]DOR: diagnostic odds ratio.

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

[c]UCPD: unsupervised correlation preserving discretization.

Table 14. Results of Experiment 4a using the differential DOR[a] and the nonoverlapping CI (keeping all pattern extensions).

Classifier, discretization, and pattern support (%)	Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[b]
J48							
Expert							
10	13	42	3.23	94.18	44.83	84.95	0.672
8	13	55	4.23	95.50	55.17	87.96	0.743
6	17	78	4.59	97.88	47.13	88.39	0.711
UCPD[c]							
16	20	74	3.7	94.97	50.57	86.67	0.761
14	7	28	4	98.41	58.62	90.97	0.804
12	12	50	4.17	100.00	65.52	93.55	0.820
JRIP							
Expert							
10	4	13	3.25	93.12	29.89	81.29	0.622
8	12	40	3.33	94.44	29.89	82.37	0.625
6	20	74	3.7	91.80	39.08	81.94	0.668
UCPD							
16	7	24	3.43	94.44	27.59	81.94	0.632
14	6	23	3.83	97.35	32.18	85.16	0.653
12	16	63	3.94	98.68	59.77	91.40	0.795

[a]DOR: diagnostic odds ratio.

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

[c]UCPD: unsupervised correlation preserving discretization.

Discussion

Principal Findings

We have proposed different ways of using the DOR as a single indicator of diagnostic performance, by carrying out a classification of the survival of patients in an ICBU by studying their daily evolution using multivariate sequential patterns. We now discuss the factors that we have to consider to have a trade-off mainly between interpretability and classification performance.

In relation to interpretability, a model is more interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model. In this sense, the presented method shows 3 advantages: (1) the readability and interpretability of the content of the patterns, (2) the reduced length of the patterns, and (3) the small set of significant patterns selected to build the classifier.

Of these 3 advantages, the most direct one for the clinician is that the patterns themselves have an interpretation in the language understood by the clinician, who does not have to spend time looking for a correspondence between what he/she read in the pattern and his/her usual way of working. Moreover, the definition of the patterns provides not only static information about the patient at admission time, as it is usual, but also the evolution of the patient. For example, a pattern like $DIUR_3 < INC_0 < INC_0 < DIUR_3$ leads the clinician to the clinical factors related to the pattern: high diuresis and very low incomings during 4 different days.

For the second factor, if we study the length of the patterns eventually selected (Table 5), it will be noted that the majority of the patterns in the baseline experiment (using JEPs) and in the first experiment (using DOR) have a length of 6 items, whereas the majority of the patterns in the subsequent experiments have a length of 4 items. We can observe that the distribution of patterns by length has changed, with a larger number of shorter patterns in the last experiments, which are more difficult to use in a classifier, because they are more general. In subsequent Experiments 2-4, we have observed that, on the one hand, the classifier is less accurate. On the other hand, the shorter patterns are easier to understand, more general, and describe the population well, but simultaneously cover survivors and nonsurvivors.

Overall, these shorter patterns produce worse classification results when we use expert discretization with a JRIP classifier. On the one hand, expert discretization generally performs worse, because it is not based on a statistical or information theory that has been specifically designed for classification purposes, and on the other hand, JRIP provides the best performance in terms of the complexity of the tree structure, while J48 produces a high classification accuracy (as the authors explain in [43]).

With shorter patterns, however, it is easier to interpret the meaning of the patterns and explain their behavior.

With respect to the third factor, we could say that a model that allows us to achieve a good classification result with a low number of patterns (and consequently of items) is, therefore, preferable. In Table 4 we obtained the smallest number of patterns with Experiment 4 (using a differential DOR and the nonoverlapping of the CI). These patterns are simultaneously restricted by these 2 conditions, and as we have selected a small number of patterns, it might even be interesting to carry out a manual revision and a study of them (although that is out of the scope of this work).

The baseline experiment (using JEPs) and Experiment 3 (nonoverlapping CI of DOR) do not depend on a threshold value and we also obtain a reasonably small number of patterns. Nevertheless the threshold value that has been established in the other experiments (Experiments 1, 2, and 4) leads to changes in the number of patterns eventually selected. We have therefore made 2 variations in Experiment 1 (using DOR), by restricting the minimum DOR value that is necessary to select patterns (Table 8), signifying that we have been able to reduce significantly the appropriate number of patterns selected.

When we work with imbalanced data, as is usual in medical domains, it is necessary to highlight the correct classification of rarely occurring cases when compared with other general cases. It is consequently necessary to check the highest specificity to choose the best classification result, which in our experiments is produced by using UCPD automatic discretization with JEPs as a classical frequency-based discriminative measure. JEPs have usually been used to build accurate classifiers, while UCPD exploits the underlying correlation structure in the data so as to obtain the discrete intervals and ensure that the inherent correlations are preserved.

Moreover, we have generally shown that this automatic discretization performs better classifications than expert discretization. But clinicians prefer to use a reference range discretization for laboratory and physiologic values. This signifies that, for example, they prefer to use the interval (7.35, 7.45) as a normal value for *PH*, as it is usually managed in medicine. The interpretability of the classification results by using expert discretization is, therefore, a prevailing factor in our choice. A summary of the principal results of the experiments using only expert discretization is shown in Table 15.

Table 15. Comparison of experimental results with the highest specificity using expert discretization.

Experiment, classifier, and pattern support (%)		Number of patterns	Total length (items)	Average length (items/pattern)	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC[a]
JEPs[b]								
J48	8	17	84	4.94	100.00	56.32	91.83	0.782
JRIP	8	15	79	5.27	100.00	58.62	92.26	0.777
1b: DOR[c]								
J48	8	17	84	4.94	94.18	55.17	86.88	0.767
JRIP	8	14	67	4.79	95.50	62.07	89.25	0.801
2b: Differential DOR								
J48	8	21	88	4.19	87.57	62.07	82.80	0.783
JRIP	6	8	29	3.62	91.53	31.03	80.22	0.623
3b: Nonoverlapping DOR								
J48	8	16	77	4.81	94.71	58.62	87.96	0.739
JRIP	6	12	51	4.25	95.77	52.87	87.74	0.747
4b: Differential + nonoverlapping DOR								
J48	8	13	55	4.23	96.30	57.47	89.03	0.770
JRIP	6	15	57	3.8	92.59	29.89	80.86	0.626

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

[b]JEP: Jumping Emerging Pattern.

[c]DOR: diagnostic odds ratio.

If we therefore consider only expert discretization, the best classification result is achieved in Experiment 1b (using DOR), with a specificity of 62.07% and an AUC value of 0.801 (Table 8). In this experiment we simultaneously obtained patterns found in both the survivors and the nonsurvivors based on only the DOR value of each pattern.

The classification model that is easiest to comprehend and has high specificity requires only 5 patterns (with a total length of 20 items) and is achieved with UCPD and a JRIP classifier in Experiment 2b (differential DOR) using beam search for the best pattern. It obtains a specificity of 56.32% and an AUC value of 0.767 (Table 10).

If we take into consideration only expert discretization, with a J48 classifier we need at least 13 patterns (with a total length of 55 items) to obtain a specificity of 57.47% and an AUC value of 0.770 (Table 13) in Experiment 4b (using a differential and a nonoverlapping DOR).

Conclusions

In this research, we have developed a model to predict the survival of patients by considering 2 aspects: the relevance of the temporal evolution of the patients as part of the model and an interpretable model for the physicians. We have achieved these aspects by (1) using the multivariate sequential patterns used in classification models that can be easily understood by experts, (2) using a reduced number of patterns, and (3) using a language that is well known by clinicians with regard to both the discretization of values and measures of interest of the patterns.

The main contribution of this work is the proposal and evaluation of 4 ways in which to employ DOR to reduce the number of patterns and to select only the most discriminative ones, because pattern explosion is a principal problem in pattern-based classifiers. We have compared the 4 proposals with a baseline experiment using JEPs. This is, to the best of our knowledge, the first time that some of these approaches have been proposed and compared in scientific literature.

With regard to the number of patterns, the best option is that of using both a differential and a nonoverlapping DOR (as in Experiment 4). As we have increased the restrictions applied, we have significantly reduced the number of patterns, thus attaining more general, simple, and interesting patterns. With expert discretization and 10% support, there are, for example, only 198 patterns (using beam search for best pattern), and, very interestingly, these patterns cover all the patients who did not survive. Despite not being within the scope of this paper, it would be interesting for a clinician to carry out a manual interpretation of these patterns.

This experiment provides the second contribution of this paper, because we have shown that beam search with the DOR could be used in the algorithm to extract sequential patterns for classification rather than using a traditional algorithm for sequential pattern mining.

Despite the efforts made to reduce the amount and the length of patterns in Experiments 2-4, in which we have compared each pattern with its extensions, the classifier built is less accurate. The shorter patterns are easier to understand, more general, and describe the population well, but simultaneously cover survivors and nonsurvivors.

With regard to accuracy, the best classification results are, not surprisingly, produced using JEPs along with UCPD. JEPs have been extensively used to build accurate classifiers and produce better results when we use a discretization based on statistical or information theory that is specifically intended for classification. Nevertheless, we require interpretable patterns that are easy for the clinician to understand, and must therefore use a reference range discretization created by an expert. If we consider only expert discretization, the highest specificity is attained using only the DOR to select the patterns (as in Experiment 1; Table 15).

With regard to interpretability, we can observe that discretization has a great impact on classification performance at the expense of interpretability, because more and longer patterns are required. With UCPD, we require only 5 patterns (with a total length of 20 items) to build a rule set and to obtain 56.32% specificity when we use the differential DOR (see Experiment 2). With expert discretization, we need at least 13 patterns (with a total length of 55 items) to obtain a specificity of 57.47% using both a differential and a nonoverlapping DOR to select the patterns (see Experiment 4).

Our future research will consist of exploring domain-based measures to evaluate clinical patterns or to reduce the number of patterns in postprocessing to an even greater extent. In this respect, we intend to investigate more specific properties, such as closed, maximal, or minimal patterns as a trade-off between improving classification performance and not losing information or representativeness of the population. The researchers additionally intend to explore other measures and search strategies that could be integrated into new algorithms.

Acknowledgments

This work was partially funded by the SITSUS project (Ref: RTI2018-094832-B-I00), the CONFAINCE project (Ref: PID2021-122194OB-I00), supported by the Spanish Ministry of Science and Innovation the Spanish Agency for Research (MCIN/AEI/10.13039/501100011033) and, as appropriate, by ERDF A way of making Europe.

Conflicts of Interest

None declared. This work does not relate to the employment of AG at Amazon.

References

1. Batal I, Fradkin D, Harrison J, Moerchen F, Hauskrecht M. Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data. : ACM Press; 2012 Presented at: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012; August 12-16; Beijing, China p. 280-288 URL: http://europepmc.org/abstract/MED/25937993 [doi: 10.1145/2339530.2339578]

2. Bringmann B, Nijssen S, Zimmermann A. Pattern-Based Classification: A Unifying Perspective. 2009 Presented at: From Local Patterns to Global Models: Proceedings of the ECML/PKDD-09 Workshop (LeGo-09); September 7-11; Bled, Slovenia p. 36-50 URL: http://arxiv.org/abs/1111.6191
3. Fan H. Efficient Mining of Interesting Emerging Patterns and Their Effective Use in Classification (PhD thesis). The Department of Computer Science and Software Engineering, University of Melbourne. 2004. URL: http://hdl.handle.net/11343/38912 [accessed 2022-07-25]
4. Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. Data Min Knowl Disc 2007 Jan 27;15(1):55-86. [doi: 10.1007/s10618-006-0059-1]
5. He Z, Gu F, Zhao C, Liu X, Wu J, Wang J. Conditional discriminative pattern mining: Concepts and algorithms. Information Sciences 2017 Jan;375:1-15. [doi: 10.1016/j.ins.2016.09.047]
6. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. Journal of Clinical Epidemiology 2003 Nov;56(11):1129-1135. [doi: 10.1016/s0895-4356(03)00177-x]
7. Agrawal R, Srikant R. Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering. New York, NY: IEEE; 1995 Presented at: Eleventh International Conference on Data Engineering; Taipei, Taiwan; March 6-10, 1995 p. 3-14. [doi: 10.1109/icde.1995.380415]
8. Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. In: Apers P, Bouzeghoub M, Gardarin G, editors. Advances in Database Technology — EDBT '96. Berlin, Heidelberg: Springer; 1996:1-17.
9. Zaki MJ. SPADE: an efficient algorithm for mining frequent sequences. Machine Learning 2001;42(1/2):31-60. [doi: 10.1007/3-540-45357-1_32]
10. Jian Pei, Jiawei Han, Mortazavi-Asl B, Jianyong Wang, Pinto H, Qiming Chen, Mei-Chun Hsu. Mining sequential patterns by pattern-growth: the PrefixSpan approach. IEEE Trans. Knowl. Data Eng 2004 Nov;16(11):1424-1440. [doi: 10.1109/tkde.2004.77]
11. Gan W, Lin JC, Fournier-Viger P, Chao H, Yu PS. A Survey of Parallel Sequential Pattern Mining. ACM Trans. Knowl. Discov. Data 2019 Jul 17;13(3):1-34. [doi: 10.1145/3314107]
12. Li W, Han J, Pei J. CMAR: accurate and efficient classification based on multiple class-association rules. In: IEEE Xplore. New York, NY: IEEE; 2001 Presented at: 2001 IEEE International Conference on Data Mining; August 7, 2002; San Jose, CA p. 369-376. [doi: 10.1109/icdm.2001.989541]
13. Nofal M, Bani-Ahmad S. Classification Based on Association-Rule Mining Techniquese a General Survey and Empirical Comparative Evaluation. Ubiquitous Computing and Communication Journal 2010;5(3):9-17 [FREE Full text]
14. Xing Z, Pei J, Keogh E. A brief survey on sequence classification. SIGKDD Explor. Newsl 2010 Nov 09;12(1):40-48. [doi: 10.1145/1882471.1882478]
15. Hu B, Chen Y, Keogh E. Time Series Classification under More Realistic Assumptions. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2013 Presented at: Proceedings of the 2013 SIAM International Conference on Data Mining; May 2-4, 2013; Texas, USA p. 578-586. [doi: 10.1137/1.9781611972832.64]
16. Drezewski R, Dziuban G, Hernik L, Paczek M. Comparison of data mining techniques for Money Laundering Detection System. New York, NY: IEEE; 2015 Presented at: 2015 International Conference on Science in Information Technology (ICSITech); October 27-28, 2015; Yogyakarta, Indonesia p. 5-10. [doi: 10.1109/icsitech.2015.7407767]
17. Lesh N, Zaki M, Ogihara M. Mining features for sequence classification. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 99. New York, NY: ACM; 1999 Presented at: KDD99: The First Annual International Conference on Knowledge Discovery in Data; August 15-18, 1999; San Diego, CA p. 342-346. [doi: 10.1145/312129.312275]
18. Tseng VSM, Lee CH. CBS: A new classification method by using sequential patterns. : Society for Industrial and Applied Mathematics; 2005 Presented at: 2005 SIAM International Conference on Data Mining (SDM 2005); April 21-23, 2005; Newport Beach, CA p. 596-600. [doi: 10.1137/1.9781611972757.68]
19. Jiménez F, Sanchez G, Juarez JM. Multi-objective evolutionary algorithms for fuzzy classification in survival prediction. Artif Intell Med 2014 Mar;60(3):197-219. [doi: 10.1016/j.artmed.2013.12.006] [Medline: 24525210]
20. Geng L, Hamilton HJ. Interestingness measures for data mining. ACM Comput. Surv 2006 Sep 30;38(3):9. [doi: 10.1145/1132960.1132963]
21. Li J, Fu AWC, He H, Chen J, Jin H, McAullay D, et al. Mining risk patterns in medical data. In: KDD '05: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York, NY: ACM; 2005 Presented at: KDD05: The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 21-24, 2005; Chicago, IL p. 770-775. [doi: 10.1145/1081870.1081971]
22. Li J, Fu AW, Fahey P. Efficient discovery of risk patterns in medical data. Artif Intell Med 2009 Jan;45(1):77-89. [doi: 10.1016/j.artmed.2008.07.008] [Medline: 18783927]
23. Wu S, Zhao Y, Zhang H, Zhang C, Cao L, Bohlscheid H. Debt Detection in Social Security by Adaptive Sequence Classification. In: Lecture Notes in Computer Science. Vol 5914 LNAI. Berlin Heidelberg: Karagiannis D, Jin Z. eds. Springer; 2009 Presented at: Knowledge Science, Engineering and Management. KSEM 2009; november 25-27; Vienna, Austria p. 192-203. [doi: 10.1007/978-3-642-10488-6_21]

24. Heierman E, Youngblood M, Cook D. Mining temporal sequences to discover interesting patterns. 2004 Presented at: Third International Workshop on Mining Temporal and Sequential Data (TDM-04); August 22, 2004; Seattle, WA.
25. Petitjean F, Li T, Tatti N, Webb GI. Skopus: Mining top-k sequential patterns under leverage. Data Min Knowl Disc 2016 Jun 14;30(5):1086-1111. [doi: 10.1007/s10618-016-0467-9]
26. Li I, Huang J, Liao I, Lin J. A sequence classification model based on pattern coverage rate. In: Lecture Notes in Computer Science, vol 7861. Springer. Berlin, Heidelberg, Germany: Springer; 2013 Presented at: Grid and Pervasive Computing: GPC 2013; May 9-11; Seoul, Korea p. 737-745. [doi: 10.1007/978-3-642-38027-3_81]
27. Toma T, Abu-Hanna A, Bosman R. Discovery and integration of univariate patterns from daily individual organ-failure scores for intensive care mortality prediction. Artif Intell Med 2008 May;43(1):47-60. [doi: 10.1016/j.artmed.2008.01.002] [Medline: 18394871]
28. Toma T, Bosman R, Siebes A, Peek N, Abu-Hanna A. Learning predictive models that use pattern discovery--a bootstrap evaluative approach applied in organ functioning sequences. J Biomed Inform 2010 Aug;43(4):578-586 [FREE Full text] [doi: 10.1016/j.jbi.2010.03.004] [Medline: 20332034]
29. Ghosh S. Multivariate Sequential Contrast Pattern Mining and Prediction Models for Critical Care Clinical Informatics (Thesis). OPUS.: University of Technology Sydney; 2017. URL: http://hdl.handle.net/10453/123204 [accessed 2022-07-25]
30. Sheppard N, Hemington-Gorse S, Shelley O, Philp B, Dziewulski P. Prognostic scoring systems in burns: a review. Burns 2011 Dec;37(8):1288-1295. [doi: 10.1016/j.burns.2011.07.017] [Medline: 21940104]
31. Casanova IJ, Campos M, Juarez JM, Fernandez-Fernandez-Arroyo A, Lorente JA. Using Multivariate Sequential Patterns to Improve Survival Prediction in Intensive Care Burn Unit. In: Lecture Notes in Computer Science, vol 9105. Cham, Switzerland: Springer; 2015 Presented at: AIME 2015: Artificial Intelligence in Medicine; June 17-20; Pavia, Italy p. 277-286. [doi: 10.1007/978-3-319-19551-3_36]
32. Allen J. Maintaining Knowledge about Temporal Intervals. Readings in Qualitative Reasoning About Physical Systems 2013;11(26):361-372. [doi: 10.1016/b978-1-4832-1447-4.50033-x]
33. Gomariz A. Techniques for the Discovery of Temporal Patterns (PhD Thesis). University of Murcia (Spain), University of Antwerp (Belgium). 2014. URL: http://hdl.handle.net/10201/38109 [accessed 2022-07-25]
34. Dong G, Li J. Efficient mining of emerging patterns. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '99. New York, NY: ACM; 1999 Presented at: KDD99: The First Annual International Conference on Knowledge Discovery in Data; August 15-18, 1999; San Diego, CA p. 43-52. [doi: 10.1145/312129.312191]
35. Dong G, Li J, Zhang X. Discovering Jumping Emerging Patterns and Experiments on Real Data sets. 1999 Jul Presented at: 9th International Database Conference on Heterogeneous and Internet Databases (IDC); July 15-17, 1999; Hong Kong p. 15-17 URL: http://corescholar.libraries.wright.edu/knoesis/402
36. Li J, Dong G, Ramamohanarao K. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. Knowledge and Information Systems 2001 May;3(2):131-145. [doi: 10.1007/pl00011662]
37. Dong G, Zhang X, Wong L, Li J. CAEP: Classification by aggregating emerging patterns. In: Lecture Notes in Computer Science. Vol 1721.: Springer Berlin Heidelberg; 1999 Presented at: International Conference on Discovery Science (DS 1999); December, 6-8; Tokyo, Japan p. 30-42. [doi: 10.1007/3-540-46846-3_4]
38. Li J, Liu J, Toivonen H, Satou K, Sun Y, Sun B. Discovering statistically non-redundant subgroups. Knowledge-Based Systems 2014 Sep;67:315-327. [doi: 10.1016/j.knosys.2014.04.030]
39. Toti G, Vilalta R, Lindner P, Price D. Effect of the Definition of Non-Exposed Population in Risk Pattern Mining. 2016 Jan Presented at: In 5th Workshop on Data Mining for Medicine and Healthcare; May 7, 2016; Miami, FL p. 5.
40. Toti G, Vilalta R, Lindner P, Lefer B, Macias C, Price D. Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. Artif Intell Med 2016 Nov;74:44-52. [doi: 10.1016/j.artmed.2016.11.003] [Medline: 27964802]
41. Casanova IJ, Campos M, Juarez JM, Fernandez-Fernandez-Arroyo A, Lorente JA. Impact of time series discretization on intensive care burn unit survival classification. Prog Artif Intell 2017 Jun 8;7(1):41-53. [doi: 10.1007/s13748-017-0130-8]
42. Daud NR, Corne DW. Human readable rule induction in medical data mining. In: Lecture Notes in Electrical Engineering. Vol 27 LNEE. Boston, MA: Springer; 2009 Presented at: Proceedings of the European Computing Conference; June 26 - 28, 2009; Tbilisi Georgia p. 787-798. [doi: 10.1007/978-0-387-84814-3_79]
43. Mohamed WNHW, Salleh MNM, Omar AH. A comparative study of Reduced Error Pruning method in decision tree algorithms. : IEEE; 2012 Presented at: 2012 IEEE International Conference on Control System Computing and Engineering, ICCSCE 2012; 23 – 25 November 2012; Penang, Malaysia p. 392-397. [doi: 10.1109/iccsce.2012.6487177]
44. Liu X, Wu J, Gu F, Wang J, He Z. Discriminative pattern mining and its applications in bioinformatics. Brief Bioinform 2015 Sep 28;16(5):884-900. [doi: 10.1093/bib/bbu042] [Medline: 25433466]

Abbreviations

AUC: area under the receiver operating characteristic curve
CBA: Classification Based on Associations

CBS: Classify-By-Sequence
CMAR: Classification Based on Multiple Association Rules
CPAR: Classification Based on Predictive Association Rules
DOR: diagnostic odds ratio
EP: emerging pattern
FN: false negative
FP: false positive
ICBU: intensive care burn unit
JEP: Jumping Emerging Pattern
MMAC: Multi-class, Multi-label Associative Classification
RIPPER: Repeated Incremental Pruning to Produce Error Reduction
SOFA: Sequential Organ Failure Assessment
TN: true negative
TP: true positive
UCPD: unsupervised correlation preserving discretization

Edited by C Lovis; submitted 22.07.21; peer-reviewed by D Hu, M Nuutinen, A Arbabisarjou; comments to author 02.01.22; revised version received 26.02.22; accepted 27.03.22; published 10.08.22.

Please cite as:
Casanova IJ, Campos M, Juarez JM, Gomariz A, Lorente-Ros M, Lorente JA
Using the Diagnostic Odds Ratio to Select Patterns to Build an Interpretable Pattern-Based Classifier in a Clinical Domain: Multivariate Sequential Pattern Mining Study
JMIR Med Inform 2022;10(8):e32319
URL: https://medinform.jmir.org/2022/8/e32319
doi:10.2196/32319
PMID:35947437

Viewpoint

Harnessing the Electronic Health Care Record to Optimize Patient Safety in Primary Care: Framework for Evaluating e–Safety-Netting Tools

Georgia Bell Black[1*], PhD; Afsana Bhuiya[2*], MRCGP; Claire Friedemann Smith[3*], PhD; Yasemin Hirst[1*], PhD; Brian David Nicholson[3*], MRCGP, DPhil

[1]Department of Applied Health Research, University College London, London, United Kingdom
[2]North Central London Cancer Alliance, London, United Kingdom
[3]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom
[*]all authors contributed equally

Corresponding Author:
Georgia Bell Black, PhD
Department of Applied Health Research
University College London
1-19 Torrington Place
London, WC1E 7HB
United Kingdom
Phone: 44 2031083157
Email: g.black@ucl.ac.uk

Abstract

The management of diagnostic uncertainty is part of every primary care physician's role. e–Safety-netting tools help health care professionals to manage diagnostic uncertainty. Using software in addition to verbal or paper based safety-netting methods could make diagnostic delays and errors less likely. There are an increasing number of software products that have been identified as e–safety-netting tools, particularly since the start of the COVID-19 pandemic. e–Safety-netting tools can have a variety of functions, such as sending clinician alerts, facilitating administrative tasking, providing decision support, and sending reminder text messages to patients. However, these tools have not been evaluated by using robust research designs for patient safety interventions. We present an emergent framework of criteria for effective e–safety-netting tools that can be used to support the development of software. The framework is based on validated frameworks for electronic health record development and patient safety. There are currently no tools available that meet all of the criteria in the framework. We hope that the framework will stimulate clinical and public conversations about e–safety-netting tools. In the future, a validated framework would drive audits and improvements. We outline key areas for future research both in primary care and within integrated care systems.

(JMIR Med Inform 2022;10(8):e35726) doi:10.2196/35726

KEYWORDS

primary care; patient safety; electronic health record; safety; optimize; framework; evaluation; tool; diagnostic; uncertainty; management; netting; software; criteria

Introduction

Safety-netting was first formally defined in the mid-1980s by Neighbour [1] and has since come to be viewed as a best practice for managing diagnostic uncertainty [2]. This is particularly relevant to primary care, wherein clinicians hold responsibility for weighing up the costs, risks, and benefits of monitoring symptoms against those of ordering tests, investigations, and referrals for further care. Safety-netting includes verbally advising to patients to practice self-care, monitor symptoms, or seek further advice if their symptoms have not resolved. Safety-netting is part of many primary care presentations, given the high volume of patients with undifferentiated nonspecific symptoms. For these patients, serious disease is a rare but important component of a differential diagnosis [3,4].

Several studies have highlighted the importance of recording safety-netting advice in patient records [5-7]. Examples of such advice include ensuring that at-risk patients are monitored, providing a reminder of the advice, facilitating the continuity of care, and maintaining a medical-legal record. Despite their

importance, safety-netting advice is not often recorded in medical notes [8]. There have been calls to improve the recording of safety-netting to facilitate follow-up and monitoring. More recently, commercial e–safety-netting tools have been developed to assist health care professionals in managing diagnostic uncertainty [9-11]. These tools may be integrated within the electronic health record (EHR) or provided by a third-party application.

The aim of this paper is to consider how e–safety-netting tools need to be developed in order to improve diagnostic safety in primary care. We also outline an emergent framework of criteria for e–safety-netting tools that can be used to facilitate evaluation and outcome measurement [12].

Safety and Safety-Netting in Primary Care

The management of diagnostic uncertainty in primary care is a part of every primary care physician's role [13]. Safety-netting mitigates the risks associated with some techniques, thereby allowing physicians to manage diagnostic uncertainty. For example, the safe use of the "test of time" allows for the expected progression of a primary care physician's initial diagnosis to be observed. Safety-netting increases safety by providing patients with information about concerning symptoms and what to do if they arise [8-10,14]. Signposting to other sources of information or to other services (eg, out-of-hours services) is also a common component of safety-netting [10]. Effective safety-netting is important, since it can have implications for a patient's outcomes by preventing misdiagnoses, complications, and delayed referrals [3,15]. It may also have workload implications by safely reducing the number of unnecessary reconsultations [15,16]. Historically, safety-netting processes have been the focus of quality improvement within the cancer clinical and research community, ranging from national strategy documents to local system providers. In health care policy and research, safety-netting has been particularly identified as a tool for facilitating the timely diagnosis of cancer [17-19].

Effective safety-netting results in patient self-care, patients' recognition of the need for and their prompt seeking of further medical attention, and the timely follow-up of patients [19]. High-quality safety-netting requires clinicians to understand a patient's information needs, the reasons for safety-netting advice, and the expected clinical course of a condition. Breakdowns in safety-netting communication could occur through the omission of information, by providing information in a way that is not easily understood or remembered, or by failing to address patient concerns [19,20]. Inconsistencies in safety-netting delivery may also harm how advice is perceived and adhered to by patients [9]. Therefore, e–safety-netting tools have a particular role in supporting clinicians' and patients' communication, information provision, knowledge, and memory.

Harnessing the EHR: e–Safety-Netting Tools—How Might They Solve Some of the Problems Above?

EHRs have been mandated for many years in primary care. These systems have been developed to capture clinical information in a way that is clinically relevant and user-friendly. EHR providers regularly update their systems to ensure that users are able to record and retrieve information easily. Over time, EHR systems have built capabilities for supporting wider functionalities, so that clinicians and managers can better support their patient populations. Although safety-netting is embedded into national health care strategies and policies, it is unclear who holds responsibility for it and how it should work [18,21]. Safety-netting is no longer considered solely as a doctor-patient interaction but as a responsibility of the "system," which should provide robust safety-netting protocols within the EHR [22]. As patients move through the multiple clinical contacts that lead up to a diagnosis, the increased specification of the safety-netting process could reduce the amount of errors in the diagnostic process [2].

e–Safety-netting tools can be integrated into the EHR or be provided by a separate piece of software. Typical functions include, for example, clinician alerts, administrative tasking, templates for standardized codes, tracking dashboards, and additional support (eg, prepopulated referral forms). The tools may support clinicians by tracking patients over a defined time interval, providing templates to guide consultations, or suggesting appropriate referral pathways [23-25]. They may also support patients by sending them trigger text reminders. Using e–safety-netting software in addition to verbal or paper-based safety-netting methods could reduce the amount of diagnostic errors and delays. This could also make improvement easier via the provision of better audit data about safety-netting. The COVID-19 pandemic has driven a surge of new e–safety-netting tools. However, these have not been evaluated by using robust research designs for patient safety interventions [12]. The variations in designs and functions suggest a lack of clarity with regard to how the tools should prevent diagnostic errors and delays.

What Safety-Netting Failures Could Be Prevented by an e–Safety-Netting Tool?

There is a lack of robust evidence suggesting whether e–safety-netting tools prevent the types of errors that they are designed to prevent. We found 2 evaluation reports of C the Signs (C the Signs Limited)—a software tool for supporting cancer decision-making and management that has been commissioned in various locations in England, United Kingdom. One evaluation found increased cancer detection rates for clinical commissioning groups, who had implemented the tools, when compared to those for groups who had not implemented the tools [26]. However, a second, independent evaluation of C the Signs found that changes to the number of referrals were inconclusive. This report, which was titled *C the Signs evaluation: report for RM Partners* (Frontier Economics, private

report received upon request, 2021), found that there was limited evidence of improved cancer detection.

e–Safety-netting tools require substantial further development in order to reach their potential in reducing the amount of diagnostic delays and errors in primary care. In Table 1, we consider the exemplar of an urgent cancer referral pathway (ie, a primary care process in which patients with suspected cancer symptoms are expected to be seen within 2 weeks for further investigations). We give details about the typical errors and outcomes that occur and how e–safety-netting tools may be developed to prevent this [27]. We also indicate whether certain functions have already been developed in prominent e–safety-netting tools that are currently available [23,25,28,29]. Future e–safety-netting tools could explore other potential process errors associated with safety-netting, such as automatically generating an alert if a patient has a number of attendances within a short span of time [30].

Table 1. Types of errors that may be mitigated by an e–safety-netting tool. We use the exemplar of an urgent cancer referral pathway.

Setting	Clinical action	Error	Outcome	Role of the e–safety-netting tool	Currently available
Doctor-patient encounter	Primary care physician is unsure whether to refer a patient with abdominal pain to specialist	Physician decides not to investigate further, as they are not aware of clinical guidelines	Delay in investigation or patient referral	Clinical presentation prompts physician to review clinical decision support tool, which reminds primary care physician of the clinical guidelines	Partially
Doctor-patient encounter	Patient visits physician multiple times for the same persistent problem	Physician does not realize that the patient has visited multiple times	Delays in taking action despite a persistent problem	Tool identifies the repeat pattern from coded data and alerts physician	No
After a consultation	Patient with low-risk symptoms is actively monitored	Patient does not reconsult a physician within the expected time frame	Delay in the timely review of symptoms	Tool alerts physician to any delays in the expected reconsultation time frame	Yes
Physician follow-up	Patient is given advice about the need for a suggested investigation	Patient is unclear about the timely review of results or how to obtain results	Delays in taking action after investigation findings	Trigger patient text message regarding reconsulting a physician promptly when results of the investigation are available	Partially
Practice level	Patient is sent to an urgent referral	Patient does not attend the urgent referral	No urgent review by a specialist	Tool identifies nonattendance and sends a message to the patient and primary care physician	Yes
Regional level	Patient is diagnosed with cancer through an emergency pathway	Primary care network does not use this as an opportunity for audit and improvement	Lack of system improvement	Nominated lead for network can review all cancer cases and disseminate learnings	Yes
Patient health record data	Patient with low-risk symptoms presents to primary care physician, resulting in self-care at home	Patient history, including risk factors, is not recorded or visible in health record	Physician is not aware of risk factors in the patient's history	Alert the physician to the incomplete patient record, including hidden risk factors, during the consultation	No
Patient health record—population	Patient's clinical risk percentage for a certain condition increases prominently (per the patient's coded data)	The data are not observed as a whole, and significant patterns are not established	The system does not identify the patient as one requiring further action	Alerts to practice-level team state that clinical risk has reached a specified trigger level for further action (investigations and referrals)	Yes

Establishing a Framework for What a Good e–Safety-Netting Tool Would Do

e–Safety-netting tool development may be viewed as an extension of EHR tool development. Hitherto, e–safety-netting tools have not been tested with respect to diagnostic safety. There are many frameworks and evidence bases on this topic. We synthesized the relevant parts of 3 publications in particular—(1) the World Health Organization *Technical Series on Safety in Primary Care: Diagnostic Errors*, which addresses how to improve the safety of multiple aspects of diagnostic and administrative work in primary care [31]; (2) Murphy and colleagues' [32] Safer Dx Trigger Tools Framework, which outlines good practice for the development of electronic tools to improve diagnostic safety; and (3) Vincent and Almaberti's [33] compendium of safety strategies. Some additional papers and our own knowledge of safety-netting and e–safety-netting tools were used to construct an emergent framework for e–safety-netting tool development (Table 2) [34-36]. This framework may be useful for audits, for e–safety-netting tool development and improvement, and for guiding future research.

Table 2. Emergent framework of principles for high-quality e–safety-netting tools.

e–Safety-netting principle	Details	Example
All patients registered will be e–safety-netted.	The tool supports reductions in diagnostic errors for all patients with all types of presentations, not just those who are considered at-risk patients.	The tool has automatic functions that work for all patients (eg, detecting multiple presentations or consultation patterns that might indicate that action is needed and triggering alerts).
All clinicians and primary care staff are responsible for e–safety-netting.	The tool is not reliant on sign-up but is automatically applied for every user registered on the system. The responsibilities would be configured to the users' credentials (eg, primary care physician, nurse, and receptionist).	The e–safety-netting functions are integrated into the electronic health record and cannot be switched off. Algorithms and alerts are live for every patient.
Limit burden and cognitive bias by using automatic functions, where possible.	The tool functions equally for every patient, not just those selected by the primary care professional or those on a "list."	Data capture is facilitated by standardized autofill. Patients are automatically selected for follow-up by risk stratification tools.
Support diagnostic processes before, during, and after consultations [34].	The tool supports continuous improvements in data quality and decision-making during the consultation, and it offers memory aids and alerts for both professionals and patients.	The tool notifies primary health care professionals when a patient data record is incomplete. Alerts are triggered or sent to a patient as a reminder to attend an investigation. The physician and patient are alerted when the patient has not attended an investigation, or the physician is alerted when the patient has not attended a specialist appointment.
Monitor, auto-detect, and measure pathway process errors or deviations and alert the relevant people [35].	The tool monitors all appropriate parts of the patient pathway. It automatically detects, rationalizes, and quantifies errors. It also alerts the appropriate staff member to errors of interest.	The tool automatically measures the time interval since the last consultation and agreed upon action. So, if there is delay in presentation, an alert is triggered. If the tool detects that a patient has not fulfilled the prescription, it alerts their health care professional and the patient.
Use simple processes that make it easy to access and transfer complex information.	The tool is easy to navigate, seamless with existing electronic health records, and automatically present at the point of care to support decision-making. Only 1 tool is in use within the primary care system to avoid confusion.	The tool allows for the easy transfer of information to other organizations and has simple and intuitive displays. It also allows users to access up-to-date pathways and referral criteria and has decision support functionalities.
Spread responsibilities and roles within primary care that have an overall impact on the whole patient pathway.	The tool allows the whole clinical and administrative team to use the tool with a centralized alert system, including champions or experts within the team.	There is shared responsibility for "flags" and errors within the system and thus a higher likelihood that the tool will initiate action. The tool supports a culture of shared responsibility.
Support senior leadership to optimize safety strategies within a regular quality improvement program.	The tool creates visual aggregate displays of increased errors (ie, practice dashboards) to establish normative quality standards. It has the ability to self-monitor and self-improve (ie, through artificial intelligence, it improves itself with data and feedback) [11].	The tool allows for the automatic identification of common diagnostic process errors, sends alerts for unexpected increases in error, and has control over the granularity of data.
Allow for patient interaction and feedback [36].	Patients can interact to input either their own health metrics or feedback on symptom changes. Patients can access the appropriate level of information to support themselves in managing their health. Integration with other e-consulting tools is possible.	Patients can self-report attendance to appointments and tick it off. Patients can provide feedback on changes in symptoms to trigger a follow-up appointment. Patients can record and report their weight or blood pressure.

Table 2 outlines 9 principles for e–safety-netting tools that we suggest would denote a high-quality tool. There are currently no tools available that meet all of the criteria in the framework. We hope that the framework will facilitate the development and improvement of e–safety-netting tools. It may also enable national and local audits and analyses, highlighting differences in performance and presenting potential solutions for improvement. Building on the development of new or modified e–safety-netting tools, health system leaders will need to ensure that their organizations have the necessary resources to implement them and to manage and respond to the data generated.

Discussion

We have presented a framework for structuring the development, evaluation, and implementation of e–safety-netting tools in primary care. The framework includes individual user benefits, technical features, and social aspects of use. Using this framework could support the progress of policies to facilitate the earlier diagnosis of serious diseases, such as cancer, cardiovascular disease, lung disease, diabetes, renal failure, and heart failure [21,37], and increase patient safety [32,38].

The framework is based on principles from established EHR tool development and patient safety frameworks but requires further validation through clinical and public input as well as empirical research. e–Safety-netting development via the use of this framework may require multidisciplinary applied research teams, including software developers; user experience and design; clinical knowledge; applied psychology; health services research; and epidemiological expertise.

The e–safety-netting framework proposed provides an approach to appraising existing tools and guiding e–safety-netting tool development. It would be valuable for commissioners to learn not only from existing experiences of successful adoption but also from decisions to decommission e–safety-netting tools [39]. Currently, there are few opportunities to understand the impact of each available e–safety-netting tool, as they are rarely evaluated and their functions are often updated. Policy makers should make it a condition that these tools be independently evaluated with results that are kept in a centrally held repository [40]. Evaluations would inform local adoption and allow for the alignment of these systems with health care strategies.

Patients need a robust, evidence-based system to ensure that they are monitored until their symptoms have been explained. Without this, primary care services are prone to operational failure. Operational failures (disruptions, errors, or inadequacies in the information, supplies, or equipment needed for patient care) are linked to often time-consuming compensatory actions for ensuring that patient care is coordinated and remains safe. At a time when workloads are continuing to increase in primary care and the format of clinical contacts is changing, e–safety-netting tools offer an approach to distributing the responsibility for follow-up safely among members of practice teams and to patients [41,42]. This is relevant to the development of integrated digital care records and population health management dashboards by integrated care systems [43]. There is further potential to look at the development of e–safety-netting at scale in secondary care and elsewhere [44].

There are likely to be challenges to uptake and implementation, even for tools that conform to the framework we have outlined [45]. However, e–safety-netting tools that align with the individual, social, and technical aspects of primary care work are more likely to succeed [46].

Acknowledgments

GBB is supported by a Health Foundation grant that was awarded to the University of Cambridge for The Healthcare Improvement Studies Institute. BDN is a National Institute for Health and Care Research academic clinical lecturer.

Conflicts of Interest

None declared.

References

1. Neighbour R. The Inner Consultation. New York, NY: Springer; 1987.
2. Nicholson BD, Mant D, Bankhead C. Can safety-netting improve cancer detection in patients with vague symptoms? BMJ 2016 Nov 09;355:i5515. [doi: 10.1136/bmj.i5515] [Medline: 28291732]
3. Almond S, Mant D, Thompson M. Diagnostic safety-netting. Br J Gen Pract 2009 Nov;59(568):872-874 [FREE Full text] [doi: 10.3399/bjgp09X472971] [Medline: 19861036]
4. Hirst Y, Lim AWW. Acceptability of text messages for safety netting patients with low-risk cancer symptoms: a qualitative study. Br J Gen Pract 2018 May;68(670):e333-e341 [FREE Full text] [doi: 10.3399/bjgp18X695741] [Medline: 29581127]
5. Chen C, Crowley R. Improving assessment of children with suspected respiratory tract infection in general practice. BMJ Open Qual 2019 Apr 08;8(2):e000450 [FREE Full text] [doi: 10.1136/bmjoq-2018-000450] [Medline: 31206053]
6. Bertheloot K, Deraeve P, Vermandere M, Aertgeerts B, Lemiengre M, De Sutter A, et al. How do general practitioners use 'safety netting' in acutely ill children? Eur J Gen Pract 2016;22(1):3-8. [doi: 10.3109/13814788.2015.1092516] [Medline: 26578087]
7. Bankhead C, Heneghan C, Hewitson P, Thompson M. Safety netting to improve early cancer diagnosis in primary care: development of consensus guidelines. GP Excellence for Greater Manchester. 2011. URL: https://gpexcellencegm.org.uk/wp-content/uploads/Safety-Netting-Guidance-for-GPs.pdf [accessed 2022-06-08]
8. Edwards PJ, Ridd MJ, Sanderson E, Barnes RK. Safety netting in routine primary care consultations: an observational study using video-recorded UK consultations. Br J Gen Pract 2019 Nov 28;69(689):e878-e886 [FREE Full text] [doi: 10.3399/bjgp19X706601] [Medline: 31740458]
9. Evans J, Ziebland S, MacArtney JI, Bankhead CR, Rose PW, Nicholson BD. GPs' understanding and practice of safety netting for potential cancer presentations: a qualitative study in primary care. Br J Gen Pract 2018 Jul;68(672):e505-e511 [FREE Full text] [doi: 10.3399/bjgp18X696233] [Medline: 29739779]
10. Jones CHD, Neill S, Lakhanpaul M, Roland D, Singlehurst-Mooney H, Thompson M. The safety netting behaviour of first contact clinicians: a qualitative study. BMC Fam Pract 2013 Sep 25;14:140 [FREE Full text] [doi: 10.1186/1471-2296-14-140] [Medline: 24066842]

11. Jones OT, Calanzani N, Saji S, Duffy SW, Emery J, Hamilton W, et al. Artificial intelligence techniques that may be applied to primary care data to facilitate earlier diagnosis of cancer: Systematic review. J Med Internet Res 2021 Mar 03;23(3):e23483 [FREE Full text] [doi: 10.2196/23483] [Medline: 33656443]
12. Brown C, Lilford R. Evaluating service delivery interventions to enhance patient safety. BMJ 2008 Dec 17;337:a2764. [doi: 10.1136/bmj.a2764] [Medline: 19091764]
13. Malterud K, Guassora AD, Reventlow S, Jutel A. Embracing uncertainty to advance diagnosis in general practice. Br J Gen Pract 2017 Jun;67(659):244-245 [FREE Full text] [doi: 10.3399/bjgp17X690941] [Medline: 28546389]
14. Alam R, Cheraghi-Sohi S, Panagioti M, Esmail A, Campbell S, Panagopoulou E. Managing diagnostic uncertainty in primary care: a systematic critical review. BMC Fam Pract 2017 Aug 07;18(1):79 [FREE Full text] [doi: 10.1186/s12875-017-0650-0] [Medline: 28784088]
15. Jones CHD, Neill S, Lakhanpaul M, Roland D, Singlehurst-Mooney H, Thompson M. Information needs of parents for acute childhood illness: determining 'what, how, where and when' of safety netting using a qualitative exploration with parents and clinicians. BMJ Open 2014 Jan 14;4(1):e003874 [FREE Full text] [doi: 10.1136/bmjopen-2013-003874] [Medline: 24430877]
16. To understand and improve the experience of parents and carers who need advice when a child has a fever (high temperature). Royal College of Paediatrics and Child Health. URL: https://tinyurl.com/4uuj9xjc [accessed 2022-06-08]
17. Jones D, Dunn L, Watt I, Macleod U. Safety netting for primary care: evidence from a literature review. Br J Gen Pract 2019 Jan;69(678):e70-e79 [FREE Full text] [doi: 10.3399/bjgp18X700193] [Medline: 30510099]
18. National Institute for Health and Care Excellence (NICE). 2020 Surveillance of Suspected Cancer: Recognition and Referral (NICE Guideline NG12). London: National Institute for Health and Care Excellence (UK); 2020.
19. Smith CF, Lunn H, Wong G, Nicholson BD. Optimising GPs' communication of advice to facilitate patients' self-care and prompt follow-up when the diagnosis is uncertain: a realist review of 'safety-netting' in primary care. BMJ Qual Saf. Epub ahead of print 2022 Mar 30 [FREE Full text] [doi: 10.1136/bmjqs-2021-014529] [Medline: 35354664]
20. Black G, van Os S, Renzi C, Walter F, Hamilton W, Whitaker KL. Does safety netting for lung cancer symptoms help patients to reconsult appropriately? A qualitative study. Research Square Preprint posted online on September 27, 2021. [FREE Full text] [doi: 10.21203/rs.3.rs-908030/v1]
21. Achieving world-class cancer outcomes: a strategy for England 2015-2020. NHS England. URL: https://www.england.nhs.uk/publication/achieving-world-class-cancer-outcomes-a-strategy-for-england-2015-2020/ [accessed 2022-06-08]
22. Network contract directed enhanced service (DES) early cancer diagnosis guidance. NHS England. URL: https://www.england.nhs.uk/wp-content/uploads/2020/03/network-contract-des-early-cancer-diagnosis-guidance.pdf [accessed 2022-06-08]
23. Fleming S, Nicholson BD, Bhuiya A, de Lusignan S, Hirst Y, Hobbs R, et al. CASNET2: evaluation of an electronic safety netting cancer toolkit for the primary care electronic health record: protocol for a pragmatic stepped-wedge RCT. BMJ Open 2020 Aug 24;10(8):e038562 [FREE Full text] [doi: 10.1136/bmjopen-2020-038562] [Medline: 32843517]
24. Home. accuRx. URL: https://www.accurx.com/ [accessed 2021-11-24]
25. Ardens the complete toolkit for SystmOne and EMIS web users. Ardens Healthcare Informatics. URL: https://www.ardens.org.uk/ [accessed 2022-04-25]
26. Hanlon J, Setters J, Payling M, Bakshi B, Moss J. Accelerating early identification of cancer in primary care using an artificial intelligence driven solution. York Health Economics Consortium. URL: https://www.nhsx.nhs.uk/media/documents/York_Health_Evaluation.pdf [accessed 2022-06-08]
27. Lyratzopoulos G, Vedsted P, Singh H. Understanding missed opportunities for more timely diagnosis of cancer in symptomatic patients after presentation. Br J Cancer 2015 Mar 31;112 Suppl 1(Suppl 1):S84-S91 [FREE Full text] [doi: 10.1038/bjc.2015.47] [Medline: 25734393]
28. Find cancer earlier. C the Signs. URL: https://cthesigns.co.uk/ [accessed 2022-04-25]
29. Electronic safety netting (E-SN) toolkit: adoption and adaption case studies. North Central London Cancer Alliance. URL: https://www.nclcanceralliance.nhs.uk/wp-content/uploads/2021/07/E-Safety-netting-case-study-publication.pdf [accessed 2022-06-08]
30. Thames Valley audit of patients diagnosed with cancer following an emergency presentation. NHS England. URL: https://www.england.nhs.uk/south/wp-content/uploads/sites/6/2017/01/tv-audit-emergency.pdf [accessed 2022-06-08]
31. Technical series on safer primary care: Diagnostic errors. World Health Organization. 2016. URL: https://apps.who.int/iris/rest/bitstreams/1071097/retrieve [accessed 2022-05-19]
32. Murphy DR, Meyer AN, Sittig DF, Meeks DW, Thomas EJ, Singh H. Application of electronic trigger tools to identify targets for improving diagnostic safety. BMJ Qual Saf 2019 Feb;28(2):151-159 [FREE Full text] [doi: 10.1136/bmjqs-2018-008086] [Medline: 30291180]
33. Vincent C, Amalberti R. Safer Healthcare: Strategies for the Real World. Cham, Switzerland: Springer; 2016.
34. Iqbal U, Celi LA, Li YCJ. How can artificial intelligence make medicine more preemptive? J Med Internet Res 2020 Aug 11;22(8):e17211 [FREE Full text] [doi: 10.2196/17211] [Medline: 32780024]
35. Rahimi SA, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: Systematic scoping review and critical appraisal. J Med Internet Res 2021 Sep 03;23(9):e29839 [FREE Full text] [doi: 10.2196/29839] [Medline: 34477556]

36. Wiljer D, Urowitz S, Apatu E, DeLenardo C, Eysenbach G, Harth T, Canadian Committee for Patient Accessible Health Records. Patient accessible electronic health records: exploring recommendations for successful implementation strategies. J Med Internet Res 2008 Oct 31;10(4):e34 [FREE Full text] [doi: 10.2196/jmir.1061] [Medline: 18974036]
37. NHS Long Term Plan. NHS England. URL: https://www.longtermplan.nhs.uk/ [accessed 2022-06-08]
38. Institute of Medicine. Health IT and Patient Safety: Building Safer Systems for Better Care. Washington, DC: The National Academies Press; 2012.
39. Williams I, Harlock J, Robert G, Mannion R, Brearley S, Hall K. Decommissioning health care: identifying best practice through primary and secondary research – a prospective mixed-methods study. In: Health and Social Care Delivery Research. Southampton, UK: NIHR Journals Library; 2017.
40. GP IT Futures systems and services. NHS Digital. URL: https://digital.nhs.uk/services/gp-it-futures-systems [accessed 2021-11-30]
41. Sinnott C, Georgiadis A, Park J, Dixon-Woods M. Impacts of operational failures on primary care physicians' work: A critical interpretive synthesis of the literature. Ann Fam Med 2020 Mar;18(2):159-168 [FREE Full text] [doi: 10.1370/afm.2485] [Medline: 32152021]
42. Sinnott C, Georgiadis A, Dixon-Woods M. Operational failures and how they influence the work of GPs: a qualitative study in primary care. Br J Gen Pract 2020 Oct 29;70(700):e825-e832 [FREE Full text] [doi: 10.3399/bjgp20X713009] [Medline: 32958535]
43. Sanderson M, Allen P, Osipovic D, Boiko O, Lorne C. The developing architecture of system management: Integrated care systems and sustainability and transformation partnerships. Policy Research Unit in Commissioning and the Healthcare System. 2021. URL: https://prucomm.ac.uk/assets/uploads/PRUComm_ICS_study_interim_report_feb_2021_1.pdf [accessed 2022-06-08]
44. Whear R, Thompson-Coon J, Rogers M, Abbott RA, Anderson L, Ukoumunne O, et al. Patient-initiated appointment systems for adults with chronic conditions in secondary care. Cochrane Database Syst Rev 2020 Apr 09;4(4):CD010763 [FREE Full text] [doi: 10.1002/14651858.CD010763.pub2] [Medline: 32271946]
45. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
46. Kilsdonk E, Peute LW, Jaspers MWM. Factors influencing implementation success of guideline-based clinical decision support systems: A systematic review and gaps analysis. Int J Med Inform 2017 Feb;98:56-64. [doi: 10.1016/j.ijmedinf.2016.12.001] [Medline: 28034413]

Abbreviations

EHR: electronic health record

Edited by C Lovis; submitted 15.12.21; peer-reviewed by J Walsh, K Blondon, J Holt; comments to author 22.03.22; revised version received 28.04.22; accepted 20.05.22; published 01.08.22.

Please cite as:
Black GB, Bhuiya A, Friedemann Smith C, Hirst Y, Nicholson BD
Harnessing the Electronic Health Care Record to Optimize Patient Safety in Primary Care: Framework for Evaluating e–Safety-Netting Tools
JMIR Med Inform 2022;10(8):e35726
URL: https://medinform.jmir.org/2022/8/e35726
doi:10.2196/35726
PMID:35916722

Review

Evaluation of the Clinical, Technical, and Financial Aspects of Cost-Effectiveness Analysis of Artificial Intelligence in Medicine: Scoping Review and Framework of Analysis

Jesus Gomez Rossi[1], MSc, DMD; Ben Feldberg[1], DMD; Joachim Krois[1], DPhil; Falk Schwendicke[1], MDPH, DMD, Prof Dr

Department of Oral Diagnostics, Digital Health and Health Services Research, Charité–Universitätsmedizin, Berlin, Germany

Corresponding Author:
Jesus Gomez Rossi, MSc, DMD
Department of Oral Diagnostics
Digital Health and Health Services Research
Charité–Universitätsmedizin
Aßmannshauser Str. 4-6
Berlin, 14197
Germany
Phone: 49 0049 30 450 625
Email: jesus.gomez-rossi@charite.de

Abstract

Background: Cost-effectiveness analysis of artificial intelligence (AI) in medicine demands consideration of clinical, technical, and economic aspects to generate impactful research of a novel and highly versatile technology.

Objective: We aimed to systematically scope existing literature on the cost-effectiveness of AI and to extract and summarize clinical, technical, and economic dimensions required for a comprehensive assessment.

Methods: A scoping literature review was conducted to map medical, technical, and economic aspects considered in studies on the cost-effectiveness of medical AI. Based on these, a framework for health policy analysis was developed.

Results: Among 4820 eligible studies, 13 met the inclusion criteria for our review. Internal medicine and emergency medicine were the clinical disciplines most frequently analyzed. Most of the studies included were from the United States (5/13, 39%), assessed solutions requiring market access (9/13, 69%), and proposed optimization of direct resources as the most frequent value proposition (7/13, 53%). On the other hand, technical aspects were not uniformly disclosed in the studies we analyzed. A minority of articles explicitly stated the payment mechanism assumed (5/13, 38%), while it remained unspecified in the majority (8/13, 62%) of studies.

Conclusions: Current studies on the cost-effectiveness of AI do not allow to determine if the investigated AI solutions are clinically, technically, and economically viable. Further research and improved reporting on these dimensions seem relevant to recommend and assess potential use cases for this technology.

(JMIR Med Inform 2022;10(8):e33703) doi:10.2196/33703

KEYWORDS

artificial intelligence; cost-effectiveness; systematic review; framework; health policy; research and development; cost; economics

Introduction

The most widespread definition of artificial intelligence (AI) asserts that "It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable" [1]. In the field of health care, AI is frequently referenced [2,3] as a tool [4] to improve diagnostics [5], facilitate screening [6], and optimize appointments for surgeries [7], among other use cases. Understanding these promising results requires considering AI research and development (R&D) as technically demanding and requiring consistent economic support for a long period of time. Some unique properties of AI, such as its high technical complexity and versatility of potential use cases, complicates studying AI solutions with standard cost-effectiveness analysis, which is frequent in the health care sector for pharmaceutical

interventions. This in turn complicates judging its overall value by decision-makers [8-10].

Currently, aspects to define funding decisions for the R&D of AI, such as the success rate of these enterprises and the monetization strategies to incentivize these investments, remain understudied. We believe that this problem is to a degree due to a lack of frameworks that explicitly state these exclusive dimensions of AI analysis so that reporting of solutions can be made comparable, reproducible, and useful [11].

We hence developed the following theories:

1. Without clinical relevance (a clear value proposition for a health care stakeholder), AI solutions with a valid technical component (availability and annotation of data, software component, regulatory component, etc) and with a viable monetization strategy (an appropriate payment mechanism or model) could remain irrelevant to the health care system.
2. Without fulfilling technical requirements, clinically relevant AI tools with clear and promising financial potential could remain technically unfeasible.
3. Without sufficient monetization that justifies any development and recuperates any investment, clinically and technically feasible (and even desirable) AI solutions could be economically unviable.

Previous systematic literature reviews have analyzed the available body of evidence and have concluded that very few studies assessed the economic impact of AI with sufficient methodological rigor [12]. Importantly, no review to this date has looked at AI development through a comprehensive framework that relates the economic investment [13], the clinical impact, and the technical development of the technology, considering the cost of opportunity of investing in AI projects, which is standard in the pharmaceutical industry [14]. These factors have to be taken into account to improve our understanding of AI solutions and to assess the value added to patients by incorporating these solutions [15,16].

We conducted a systematic scoping review to assess existing literature from clinical, technical, and economic perspectives. We took a scoping approach to summarize the articles included in our review and constructed a framework that facilitates the comparison of AI R&D from these 3 perspectives, according to our above-discussed theories.

Scoping reviews are replicable and systematic, and are especially suited to assess an available body of evidence and to eventually inform research and policy priorities [17-19]. Frequently, they allow an exploratory research question to be framed within the available body of evidence to expose research gaps [19]. For the summary of our scoping review, we developed a health policy framework that merged and adjusted existing frameworks to this novel technology, according to existing best practice guides for health policy analysis [20]. These included exploring a new approach for synthesis and making our assumptions explicit, logical, interrelated, and open for empirical testing while focusing on synergizing with existing methods for analyzing AI technologies.

Methods

Synthesis and Reporting

Articles included in our scoping review were analyzed according to clinical, technical, and economic dimensions relevant to AI solutions, using our framework for analysis [21]. We consider an AI solution as any algorithm capable of classifying, recommending, analyzing, or suggesting the improvement of a clinical or organizational process without previous exposure to the data analyzed. We included AI solutions developed for a specific purpose, third-party AI solutions used as software as a service, and software solutions with an AI algorithm included in the service provided.

We then designed a framework for scoping and analyzing the literature included in our scoping review. This was achieved by combining different existing frameworks according to our proposed theory and extrapolating from them [20]. We then proceeded to adjust and quantify the categories applicable to this study. We then validated it by applying it to our research methods and proceeded to assess its saturation. The aim was to assess its usability, as well as determine which components were more frequently deployed in the field.

This review was conducted following the principles modified by Levac [18,22]. Reporting follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [Multimedia Appendix 1] 2020 statement [23]. This review could not be registered in the PROSPERO database since scoping reviews are not included from October 2019. The protocol entry is available from the authors on request.

Eligibility Criteria

We included all forms of economic evaluations, reports on cost-effectiveness, and reports on the economic impact of AI solutions or AI algorithms used by any health care–related actor. Our population included patients, health care providers, insurance companies, the pharmaceutical industry, and suppliers of health care goods. Interventions would require the use of AI directly through a programing language that allows analyzing a certain database with a pre-existing open-source platform or data analysis library, as well as an integrated AI solution within a customized software. We preferred studies that compared AI against at least one comparator (ideally standard of care, ie, control), but we also assessed those without a control group, since new treatment paths or analyses may not have a clear comparator, such as in the case of fraud detection from an insurance perspective. Outcomes included in our review had a comparator of the utility, benefit, effectiveness, or cost assessed. No time limitation for the publication date was set. Our search was limited to English and German (the main languages spoken by our team).

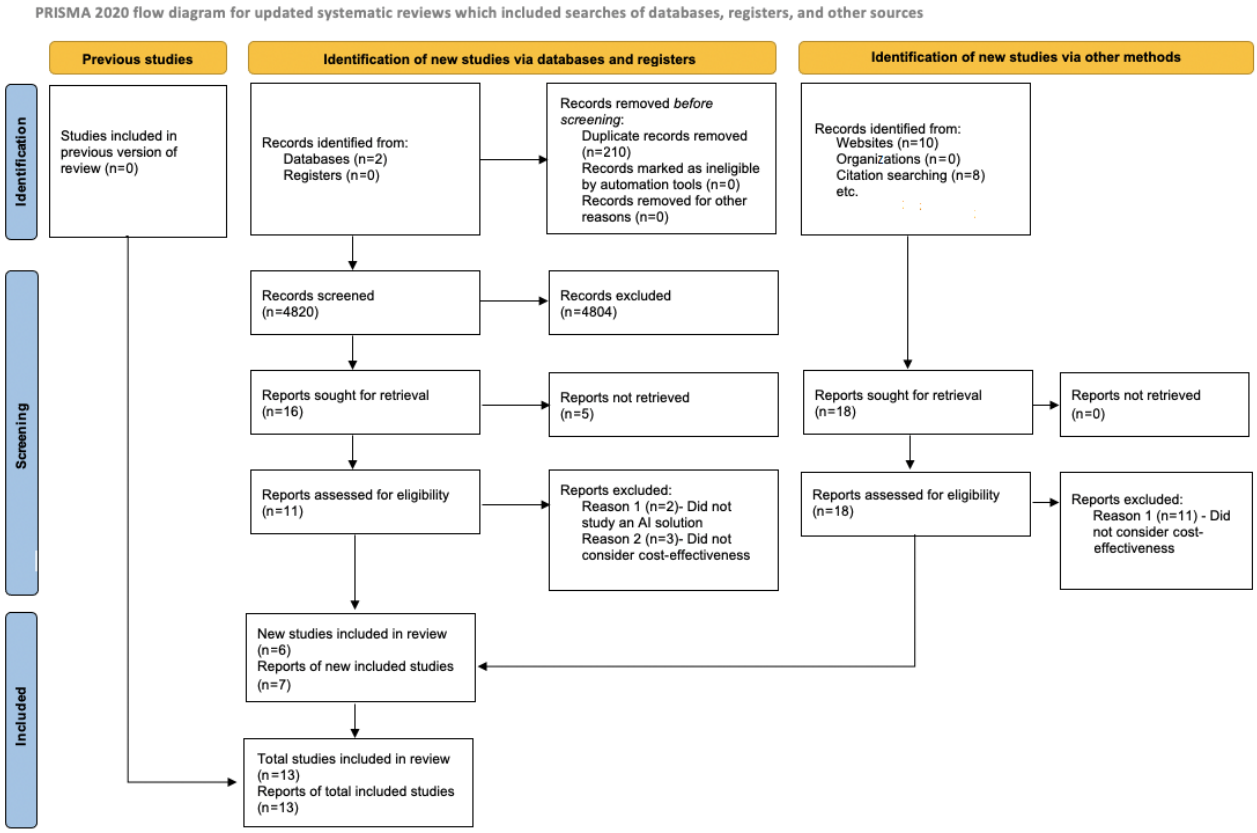
Information Sources, Search, and Study Selection

MEDLINE (via PubMed) and Embase (via Ovid) were searched for studies published until April 2021. Search strategies were adapted for each database. The following search strategy was used for PubMed: (((((((economic analysis[Title/Abstract]) OR (economic evaluation[Title/Abstract])) OR (cost-effectiveness[Title/Abstract])) OR

(monetization[Title/Abstract])) AND (artificial intelligence[Title/Abstract])) OR (Convolutional neural network[Title/Abstract])) OR (machine learning[Title/Abstract])) OR (deep learning[Title/Abstract])

Two reviewers (JGR and BF) independently screened the identified studies for eligibility. Potentially eligible studies were assessed (JGR and BF), and inclusion was decided in consensus with a third reviewer (FS). We created a list of cited sources and then proceeded to manually retrieve the sources and evaluate in full the articles for inclusion. When this second source lead to a third source and the third source met the inclusion criteria, this source was also included and classified as "citation research" in our analysis. In order to expand our scope, a hand search was performed online to look for scientific references on regulatory AI databases mentioned in the studies included, and these were included as studies identified from "websites." Details can be seen in Figure 1. The decision of not including specific medical disciplines in the search strategy was deliberate and aimed at achieving a broad inclusion of articles tailored to medical databases.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 [23] flowchart. AI: artificial intelligence.



Inclusion/Exclusion Criteria

The inclusion criteria for AI studies were as follows: (1) a solution developed using AI or any technique encompassed on it (machine learning, deep learning, etc); (2) application to any medical field/medical facility directly providing services to patients, and (3) any claim over a cost-effectiveness analysis of this technology, regardless of the methodology utilized.

The exclusion criterion was grey literature to accommodate for the lack of a risk of bias assessment in scoping reviews.

Data Collection Process and Items

Data extraction was performed by 2 reviewers independently (JG and BF) in a pilot-tested spreadsheet. Eligible studies were collected in a single spreadsheet for screening, which was performed by each reviewer independently. Studies meeting the inclusion criteria were marked by each reviewer and accepted after comparing results with those of the other reviewer. Disagreements were solved by consensus-based discussion, and if infructuous, they were solved by consulting a third reviewer (FS).

The following data were collected: year and country where the study was conducted, what outcomes were measured and how, payer's perspective assumed, comparator considered (if applicable), what benefits were measured, and what analysis was used to compare differences in the effect with the baseline case. When applicable, the following data were collected: who annotated the data for training the algorithm, how was the data set composed, image type or information type used to train the algorithm, use case assumed, AI algorithm used, and diagnostic accuracy considered (sensitivity/specificity).

Data Synthesis and Framework Construction

Our policy framework was developed according to Walt et al [20] and Hetrick et al [24] to synthesize our included articles. According to our theory, the framework for synthesis considered the following 3 dimensions: clinical aspects, technical aspects, and monetization/economic aspects. To analyze these

dimensions, 2 authors (JGR and BF) proceeded to generate 3 independent lists (1 per dimension) to include in the analysis and rank the articles by completeness, correctness, and logic consistency for the analysis of AI. Selected frameworks were then adjusted and presented to a third reviewer (FS) who assisted in maintaining the development of the tool within the scope of our research theory by having the final vote over discrepancies and presenting alternatives where evidence was not easily available.

We desisted from using tools to assess the risk of bias or assess methodological quality since scoping reviews do not aim to produce a critically appraised and synthesized result/answer to a particular question. They rather aim to provide an overview or map of the evidence. Due to this, assessing methodological limitations or risk of bias of the evidence included within a scoping review is generally not performed [17]. The PRISMA checklist for systematic scoping reviews was utilized.

The generated framework can be found in Multimedia Appendix 2.

Clinical Consideration of AI R&D

First, AI solutions were classified in the medical discipline they are supposed to be designed for (part 1a), according to the typology developed by the Association of American Medical Colleges [25] and modified to include dentistry. We did so to achieve a broader coverage of all medical disciplines contained in this framework, since dentistry may not be within the jurisdiction of medical colleges but instead dental colleges, and despite that belongs to the health care field. The categories were allergy and immunology, anesthesiology, dermatology, diagnostic radiology, emergency medicine, public health, internal medicine, medical genetics, neurology, nuclear medicine, obstetrics and gynecology, ophthalmology, pathology, pediatrics, physical medicine and rehabilitation, psychiatry, radiation oncology, surgery, urology, and dentistry.

Second, we considered the perspectives of users (1b), defined as those who use or benefit from using AI solutions. In health care, differences exist between those who use a service (reflected in this classification) and those who pay for it (analyzed in the economic considerations of this framework; 3b). The typology for this classification was extracted from the work of Sneha et al [26] who categorized value propositions in eHealth. The categories were as follows: patients, health care professionals, insurance companies, pharmaceutical companies, and vendors or suppliers.

Third, the value proposition (1c), that is, the benefit derived from using AI solutions, was analyzed. Our classification was derived from the frameworks for software and mobile health published by Gorski et al [27] and Walther et al [28], who defined value propositions for software and modified them to include AI. The categories were as follows: improved experience for users or professionals, improved data collection/curation, outsourcing of screening to another provider, improved financing, optimized direct resource utilization (medical resources utilized for the process optimized: capital or labor), optimized indirect resources (waiting times, detecting possible cancellation of appointment, etc), branding, fraud detection/quality control, risk assessment, improved recommendation of a provider/product, community building and transparency, accounting benefits, energy savings, replacement of old infrastructure (outsourcing of processes), improved data security, improved mobility, improved availability, improved ease of use, improved helpdesk quality (follow-up of cases and chatbots), facilitation of innovation, improved actualization of a service or product, and strategic flexibility (lower sunk costs).

As an "AI value proposition," we added optimization, which we defined as improved output with the same resources or reduced costs producing the same output. We acknowledge that this classification may require revision over time as AI unravels new value for users.

Finally, AI solutions were grouped (1d) according to the "EU software as medical device regulation (MDR)" from 2017 [29], using a binary categorization to assess the need for premarket approval. Because the exact determination of risks is normally independently evaluated by regulatory bodies, this classification is purposefully general to differentiate AI solutions that could be part of medical devices and affect pathways of care from those use cases that would not require market preapproval by regulatory bodies. The categories were "No" or "Class I/II/III" (needing premarket approval).

Technical Aspects of AI R&D

Developing AI solutions in health care could be particularly demanding as regulatory bodies require, in some cases, extensive testing of these products before granting them market approval. As a result, AI investors could expect high costs to enter the market and a lower success rate. It is expected that AI investors take the perspective of pharmaceutical companies and expect that the benefits from a successful digital product compensate for the high failure rate of other projects [30-34]. This is a common practice in the pharmaceutical field [35]. As a result, the framework assesses the direct R&D costs per AI solution generated that successfully enters the market (2a) and the R&D costs of the jointly developed products that do not enter the market (2b).

The direct costs per AI solution generated (2a) were divided into the 2 categories of labor and capital. We considered the following as "fixed" direct costs of AI development (paid once): data generation/acquisition, data labeling, data science, software engineering services, overheads (marketing, management, and hardware), and regulatory costs.

The costs of R&D for the pharmaceutical industry from an investor perspective (2b) comprise the costs of investing in the development of an AI solution, adjusted by the risk of failure. In this industry, previous studies have estimated the costs per new product brought to the market considering both direct and indirect (personnel and overhead) R&D costs per therapeutic product each year, adjusted by inflation to US$ using the US consumer price index [36]. Other studies have retrospectively assessed the cost of the opportunity of investing in pharmaceutical products by assessing all projects managed by a pharmaceutical company, including those that failed, and dividing total R&D costs by the costs for projects that

succeeded, to conclude the "cost of money" for these enterprises, as the real cost of capital rate has been historically 10.5% per year [14]. This allows the estimation of the required risk-free rate of return for an investor considering other investment opportunities, paid as a yearly premium [37,38]. It is likely that R&D costs of AI included in our framework would lead to significant underestimation since, to this date, there is insufficient information on the return of investment on AI in health care.

The "variable" costs or costs per good sold grow with output (2c). They were estimated based on a real-world AI solution in dentistry [39] (Multimedia Appendix 3), which considered exclusively cloud infrastructure and customer support. Although this assumption is likely to underestimate other running costs, such as improvement of the algorithm, marketing, and surveillance, among others, it seeks to make explicit that some commercial use cases of AI require a dedicated postmarket launch team that could be later added to the section "Others" of our framework.

The categories we assessed consisted exclusively of "cloud infrastructure," "customer support and quality management," "third-party products," and "other costs."

Monetization of AI

This dimension explicitly analyses AI solutions' payment mechanisms (3a) and payment model (3b). It should be noted that the potential beneficiaries and users of AI solutions are more diverse than the narrow patient perspective taken to analyze clinical outcomes in standard pharmacological products. Payment mechanisms are derived from an analysis of the value of data by Deighton et al [40]. To make possible cross-country comparisons, as well as comparisons across different use cases, we focused exclusively on payment methods irrespective of the legislation of the country we were assessing. Because there are many major differences in access to the market by different products, the results should be interpreted with caution. A certain business model dependent on a monetization scenario may likely be highly impactful, cost-effective, and profitable in one setting, but completely irrelevant, not very cost-effective, or completely illegal when extrapolated to another. Because of that, this category exclusively focuses on naming options for monetization found in the literature while remaining open to incorporating future monetization scenarios. We acknowledge that for AI developers, decisions on how and where to access a market will be conditional on a complete evaluation of a legal landscape rapidly changing and not considered in this review.

The categories in the payment mechanisms analyzed included the following: license or white labeling, one-time purchase, freemium and premium, SaaS (assuming a flat fee for each service provided), publicity, pay-for-performance, profit sharing, shared saving, bundled payment, and exclusivity contract.

An appropriate payment model is a requisite for a sound business model in digital health [41]. As discussed previously, this category helps to assess explicitly who is supposed to pay for the solution and in which contractual modality, and not who benefits from the AI solution. This category differentiates between AI solution companies focused on offering services to other companies (known as "business to business" or "B2B") and companies focused on offering the same AI services but to individual consumers ("business to consumer" or "B2C").

Risk of Bias

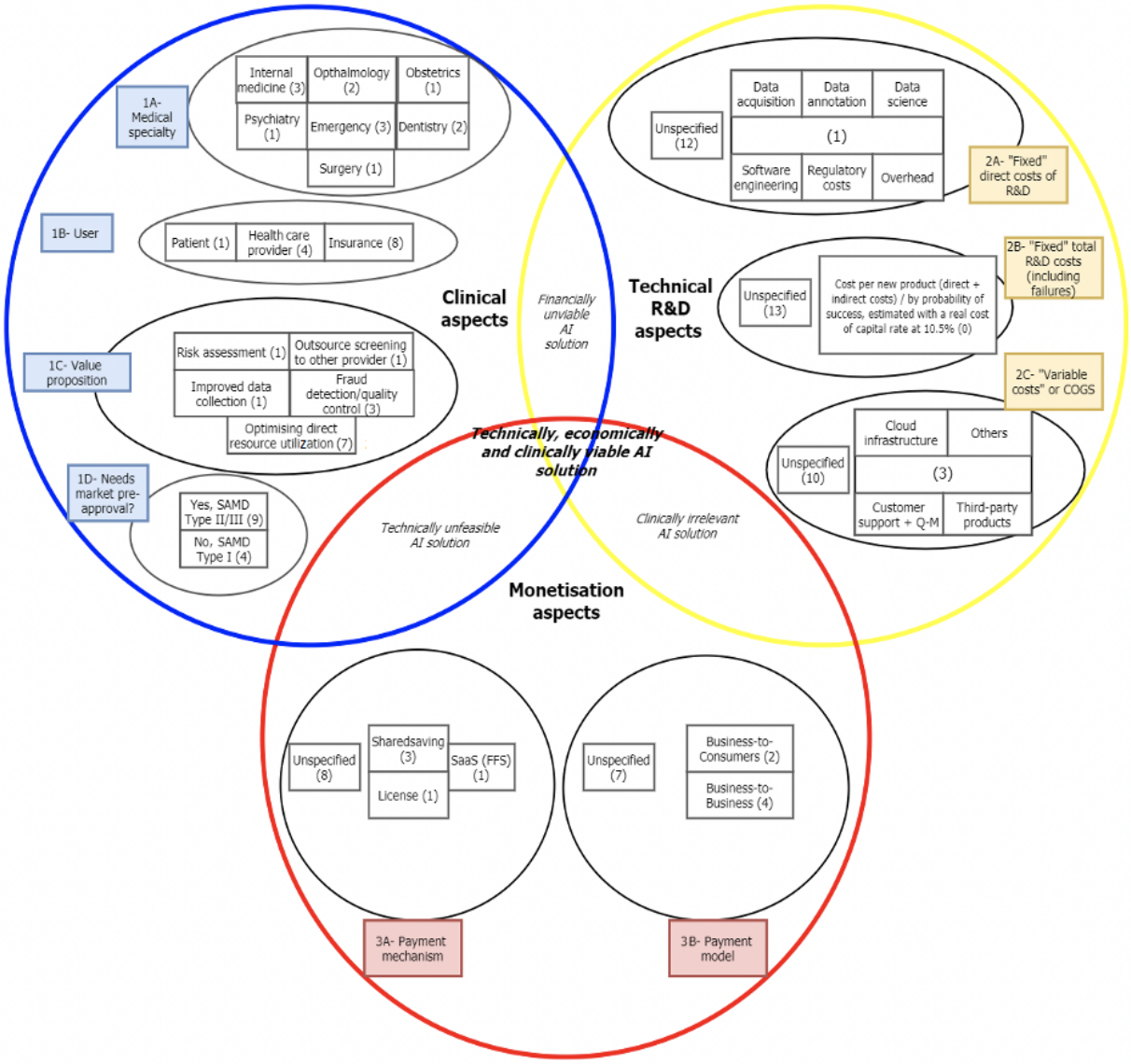
All classifications were carefully evaluated by the reviewer team (JGR, FS, and BF), and disagreements were solved by discussion. Further quantitative synthesis or evaluation of meta-biases was not feasible due to high data heterogeneity. The risk of bias or the assessment of methodological quality was not included in this review since scoping reviews do not aim to produce a critically appraised and synthesized result/answer to a particular question, as discussed [17].

Results

Included Studies and Data Description

Mapping of the identified studies is presented in Figure 2. Studies were grouped according to the clinical, technical, and economic aspects of the AI. Each article was categorized in our framework according to the pre-established categories extracted from the literature. When the information necessary for classification was not available, the corresponding AI solution was classified as "unspecified." AI solutions could only be included at 1 level.

Figure 2. Mapping of identified studies along with the developed framework. AI: artificial intelligence; COGS: costs per good sold; R&D: research and development; SAMD: software as a medical device.



We identified 4820 articles as initially eligible for our review through database screening. After screening, 16 studies were retrieved and assessed in full, and 6 were included in this review (Figure 1). Additionally, 7 other studies were included after being identified via a web search and citation search. The studies excluded did not meet the criterion of considering the dimensions of cost-effectiveness in the AI solutions they analyzed. Three systematic literature reviews met the inclusion criteria posed by our review. Two of them were conducted by a governmental body that aimed to find cost-effective therapies for diagnostic screening.

Multimedia Appendix 4 summarizes the articles meeting the inclusion criteria [39,42-53]. The studies included showed broad variability in the data used by the AI solution to generate inference and in the types of algorithms used, and frequently compared their results to the standard of care. Among the 13 studies included, 5 (39%) took place in the United States, 2 (16%) took place in Germany, and 2 (16%) took place in Canada, and Singapore, Turkey, Zambia, and the United Kingdom had 1 (7%) study each.

The majority of the studies included (9 of 13) assessed AI solutions that may require some form of premarket authorization. Internal medicine and emergency remained the most frequently studied specialties. AI solutions aimed at patients and health care providers were studied in 5 cases. The optimization of direct resource use remained the most frequent value proposition (7/13, 54%).

The technical aspects analyzed remained unaddressed to a large degree. Except for 3 (23%) studies, most of the articles analyzing the cost-effectiveness of AI solutions disregarded variable costs. Only 1 study estimated fixed costs of R&D, disregarding any reporting on opportunity costs from an investor's perspective. No study considered the costs of data acquisition or failed enterprises.

The economic aspects analyzed remained underreported to a significant degree. Furthermore, even among those studies in

which a payment model was assumed (6/13, 46%), the payment mechanism could not be identified in the use case analyzed. A minority of articles that analyzed the cost-effectiveness of AI explicitly stated the payment mechanism assumed (5/13, 38%), with a majority (8/13, 62%) insufficiently reporting the mechanism.

Discussion

Principal Findings

AI is a novel yet highly promising and versatile technology with a demonstrated capacity to undertake different tasks in the medical field with high accuracy, and unlike standard pharmaceutical products, it can help different actors of the health care system in a variety of different use cases. However, compared with other fields, standard cost-effectiveness evaluations may require adaptations, which is why this study developed a common framework for evaluation and to facilitate communication between developers, patients, doctors, and decision-makers.

The use of our framework fosters a comprehensive assessment of different dimensions of AI and makes explicit assumptions involving AI R&D, frequently overseen in previous studies. We believe that having these in mind can help to optimize research solutions where they can have the most impact by considering appropriate budgeting. Importantly, this framework could as well give both decision-makers and developers common ground to negotiate payment methods by explicitly stating the costs of development.

The analysis of our results using our framework indicates that the majority of the economic evaluations included in our study reported the clinical or organisational benefits of AI without an appropriate disclosure and justification of technical and financial aspects that substantiate these claims. It is likely that a relevant share of information and aspects is hence not fully reflected, possibly leading to biased conclusions by these studies. This seems relevant because it possibly calls for an improvement in reporting AI R&D, especially in the area of costs surrounding technical and monetization aspects to facilitate recognizing the niches where AI development will have the highest benefit for society.

Our results also invite further consideration of the setting of analysis, as regulation and market access may vary greatly and determine the economic viability of AI solutions. More transparent disclosure of clinical, technical, and economic aspects could not only generate common ground to differentiate promising projects from those excessively technically complex or clinically irrelevant, but also simplify the cooperation between AI developers, investors, clinicians, patients, and regulators.

Strengths and Limitations

First, in this review, we did not comprehensively assess the qualitative aspects of the included studies, such as their risk of bias. This limitation seemed acceptable in light of our initially planned scope, which was focused on developing a framework of analysis to gauge the comprehensiveness and completeness of existing studies. Second, although our framework could require extension with further categories of analysis and future adjustments, we believe it has succeeded in making explicit the current research gaps in the existing body of literature. Third, we acknowledge the lack of other comprehensive frameworks of analysis and limited evidence supporting our analyses, which is why this article should be considered as the start of the scientific analysis of the cost-effectiveness of AI in health care. We acknowledge that our conclusions are preliminary in a field that continues to evolve rapidly, and our results should be interpreted with caution, as future methods of analysis will have to be developed jointly with new AI solutions.

Future studies could validate or disprove the completeness of this framework and possibly work to continue and reform some of its components as AI technology continues to expand its functionality over time. Additionally, future scoping reviews could help to obtain an overview of the development of this technology over time and help to identify suitable comparisons between subfields involving AI, which could greatly facilitate generating systematic literature reviews focused on clinical effectiveness, such as meta-analyses and formal cost-effectiveness comparisons.

Conclusion

The literature reviewed in our study was sparse and did not seem comprehensive enough to draw a conclusive analysis on AI's potential to facilitate cost-effective healthcare. While some studies have showcased the positive impact of AI adoption, future research should improve reporting of the technical aspects of AI development. This seems important to achieve better comparisons of similar use cases of this novel and highly versatile technology. We believe that the adoption of the framework discussed in this study can facilitate more robust scientific analysis and better-informed conclusions on the potential of this technology.

Conflicts of Interest

FS and JK are founders of dentalxr.ai Ltd, a company developing artificial intelligence for dental diagnostics. Both authors declare no direct conflicts of interest.

Multimedia Appendix 1
PRISMA-ScR Checklist.
[DOCX File , 107 KB - medinform_v10i8e33703_app1.docx]

Multimedia Appendix 2

Clinical, technical, and economic dimensions included in our framework for analysis.
[DOCX File , 22 KB - medinform_v10i8e33703_app2.docx]

Multimedia Appendix 3
Input parameters for dentistry.
[DOCX File , 38 KB - medinform_v10i8e33703_app3.docx]

Multimedia Appendix 4
Articles included in the review.
[DOCX File , 19 KB - medinform_v10i8e33703_app4.docx]

References

1. McCarthy J. What is Artificial Intelligence? Stanford University. 2007. URL: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf [accessed 2022-06-18]
2. IDx-DR Becomes First FDA-Approved AI-Based Diagnostic for Diabetic Retinopathy. Xtalks. 2018. URL: https://xtalks.com/idx-dr-becomes-first-fda-approved-ai-based-diagnostic-for-diabetic-retinopathy-1274/ [accessed 2019-10-24]
3. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. FDA. 2018. URL: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye [accessed 2020-07-27]
4. Hwang J, Jung Y, Cho B, Heo M. An overview of deep learning in the field of dentistry. Imaging Sci Dent 2019 Mar;49(1):1-7 [FREE Full text] [doi: 10.5624/isd.2019.49.1.1] [Medline: 30941282]
5. 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 Aug 28;1(1):39 [FREE Full text] [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
6. Bellemo V, Lim G, Rim TH, Tan GSW, Cheung CY, Sadda S, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. Curr Diab Rep 2019 Jul 31;19(9):72. [doi: 10.1007/s11892-019-1189-3] [Medline: 31367962]
7. Rozario N, Rozario D. Can machine learning optimize the efficiency of the operating room in the era of COVID-19? Can J Surg 2020 Dec 01;63(6):E527-E529 [FREE Full text] [doi: 10.1503/cjs.016520] [Medline: 33180692]
8. Briggs AH, Claxton K, Sculpher MJ. Decision Modelling for Health Economic Evaluation. Oxford, United Kingdom: Oxford University Press; 2006.
9. Ramsey SD, Willke RJ, Glick H, Reed SD, Augustovski F, Jonsson B, et al. Cost-effectiveness analysis alongside clinical trials II-An ISPOR Good Research Practices Task Force report. Value Health 2015 Mar;18(2):161-172 [FREE Full text] [doi: 10.1016/j.jval.2015.02.001] [Medline: 25773551]
10. O'Brien BJ, Briggs AH. Analysis of uncertainty in health care cost-effectiveness studies: an introduction to statistical issues and methods. Stat Methods Med Res 2002 Dec 02;11(6):455-468. [doi: 10.1191/0962280202sm304ra] [Medline: 12516984]
11. Gomez Rossi J, Rojas-Perilla N, Krois J, Schwendicke F. Cost-effectiveness of artificial intelligence as a decision-support system applied to the detection and grading of melanoma, dental caries, and diabetic retinopathy. JAMA Netw Open 2022 Mar 01;5(3):e220269 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.0269] [Medline: 35289862]
12. Wolff J, Pauling J, Keck A, Baumbach J. The economic impact of artificial intelligence in health care: systematic review. J Med Internet Res 2020 Feb 20;22(2):e16866 [FREE Full text] [doi: 10.2196/16866] [Medline: 32130134]
13. COGS formula: why it matters for your business. Tradegecko. URL: https://www.tradegecko.com/blog/inventory-management/how-to-calculate-cogs [accessed 2021-04-15]
14. DiMasi JA, Grabowski HG, Hansen RW. J Health Econ 2016 May;47:20-33. [doi: 10.1016/j.jhealeco.2016.01.012] [Medline: 26928437]
15. Dismuke C. Progress in examining cost-effectiveness of AI in diabetic retinopathy screening. The Lancet Digital Health 2020 May;2(5):e212-e213 [FREE Full text] [doi: 10.1016/S2589-7500(20)30077-7]
16. 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]
17. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
18. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
19. Colquhoun HL, Levac D, O'Brien KK, Straus S, Tricco AC, Perrier L, et al. Scoping reviews: time for clarity in definition, methods, and reporting. J Clin Epidemiol 2014 Dec;67(12):1291-1294. [doi: 10.1016/j.jclinepi.2014.03.013] [Medline: 25034198]

20. Walt G, Shiffman J, Schneider H, Murray SF, Brugha R, Gilson L. 'Doing' health policy analysis: methodological and conceptual reflections and challenges. Health Policy Plan 2008 Sep;23(5):308-317 [FREE Full text] [doi: 10.1093/heapol/czn024] [Medline: 18701552]
21. Daudt HML, van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. BMC Med Res Methodol 2013 Mar 23;13:48 [FREE Full text] [doi: 10.1186/1471-2288-13-48] [Medline: 23522333]
22. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Int J Surg 2021 Apr;88:105906. [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
24. Hetrick S, Parker A, Callahan P, Purcell R. Evidence mapping: illustrating an emerging methodology to improve evidence-based practice in youth mental health. J Eval Clin Pract 2010 Dec;16(6):1025-1030. [doi: 10.1111/j.1365-2753.2008.01112.x] [Medline: 20337833]
25. Specialty Profiles. Association of American Medical Colleges. URL: https://www.aamc.org/cim/explore-options/specialty-profiles [accessed 2021-04-16]
26. Sneha S, Straub D. E-Health: Value proposition and technologies enabling collaborative healthcare. In: Proceedings of the 50th Hawaii International Conference on System Sciences. 2017 Presented at: 50th Hawaii International Conference on System Sciences; January 4-7, 2017; Hilton Waikoloa Village, Hawaii, USA. [doi: 10.24251/HICSS.2017.108]
27. Gorski I, Bram JT, Sutermaster S, Eckman M, Mehta K. Value propositions of mHealth projects. J Med Eng Technol 2016 Aug 12;40(7-8):400-421. [doi: 10.1080/03091902.2016.1213907] [Medline: 27687907]
28. Walther S, Plank A, Eymann T, Singh N, Phadke G. Success factors and value propositions of software as a service providers - A literature review and classification. AMCIS 2012 Proceedings. 2012. URL: https://aisel.aisnet.org/amcis2012/proceedings/EnterpriseSystems/1/ [accessed 2022-06-18]
29. Classification. The European Union Medical Device Regulation. URL: https://eumdr.com/classification/ [accessed 2021-04-16]
30. Bayer launches LifeHub UK focused on Artificial Intelligence to optimize data-driven drug discovery and disease diagnosis. Bayer. URL: https://media.bayer.com/baynews/baynews.nsf/id/Bayer-launches-LifeHub-UK-focused-Artificial-Intelligence-optimize-data-driven-discovery-disease [accessed 2022-06-18]
31. Artificial Intelligence. Novartis. URL: https://www.novartis.com/about/strategy/data-and-digital/artificial-intelligence [accessed 2022-06-18]
32. Partnering in a digital era. Roche. URL: https://www.roche.com/stories/partnering-in-a-digital-era [accessed 2021-04-19]
33. A New Frontier for AI: Helping Scientists Develop Potential New Medicines. Pfizer. URL: https://www.pfizer.com/news/articles/a_new_frontier_for_ai_helping_scientists_develop_potential_new_medicines [accessed 2021-04-19]
34. Bristol Myers Squibb. URL: https://www.bms.com/life-and-science/science/the-role-of-artificial-intelligence.html [accessed 2021-04-19]
35. Luehrman TA. Strategy as a Portfolio of Real Options. Harvard Business Review. 1998. URL: https://hbr.org/1998/09/strategy-as-a-portfolio-of-real-options [accessed 2021-01-15]
36. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. JAMA 2020 Mar 03;323(9):844-853 [FREE Full text] [doi: 10.1001/jama.2020.1166] [Medline: 32125404]
37. Chit A, Chit A, Papadimitropoulos M, Krahn M, Parker J, Grootendorst P. The opportunity cost of capital: development of new pharmaceuticals. Inquiry 2015 May 01;52:004695801558464 [FREE Full text] [doi: 10.1177/0046958015584641] [Medline: 25933615]
38. McNulty J, Yeh T, Schulze W, Lubatkin M. What's Your Real Cost of Capital? Harvard Business Review. 2002. URL: https://hbr.org/2002/10/whats-your-real-cost-of-capital [accessed 2021-04-19]
39. Schwendicke F, Rossi JG, Göstemeyer G, Elhennawy K, Cantu AG, Gaudin R, et al. Cost-effectiveness of artificial intelligence for proximal caries detection. J Dent Res 2021 Apr;100(4):369-376 [FREE Full text] [doi: 10.1177/0022034520972335] [Medline: 33198554]
40. Deighton JA, Johnson PA. Harvard Business School. URL: https://www.hbs.edu/faculty/Pages/item.aspx?num=48601 [accessed 2021-04-19]
41. Grustam AS, Vrijhoef H, Cordella A, Koymans R, Severens JL. Care coordination in a business-to-business and a business-to-consumer model for telemonitoring patients with chronic diseases. Int J Care Coord 2017 Dec;20(4):135-147 [FREE Full text] [doi: 10.1177/2053434517747908] [Medline: 29276610]
42. Faust O, Lei N, Chew E, Ciaccio EJ, Acharya UR. A smart service platform for cost efficient cardiac health monitoring. Int J Environ Res Public Health 2020 Aug 30;17(17):6313 [FREE Full text] [doi: 10.3390/ijerph17176313] [Medline: 32872667]

43. Huff TJ, Ludwig PE, Zuniga JM. The potential for machine learning algorithms to improve and reduce the cost of 3-dimensional printing for surgical planning. Expert Rev Med Devices 2018 May 09;15(5):349-356. [doi: 10.1080/17434440.2018.1473033] [Medline: 29723481]
44. Lachance C, Ford C. Portable Stroke Detection Devices for Patients with Stroke Symptoms: A Review of Diagnostic Accuracy and Cost-Effectiveness. Ottawa, ON, Canada: Canadian Agency for Drugs and Technologies in Health; 2019.
45. Lachance C, Walter M. Artificial Intelligence for Classification of Lung Nodules: A Review of Clinical Utility, Diagnostic Accuracy, Cost-Effectiveness, and Guidelines. Ottawa, ON, Canada: Canadian Agency for Drugs and Technologies in Health; 2020.
46. Lee HK, Jin R, Feng Y, Bain PA, Goffinet J, Baker C, et al. An analytical framework for TJR readmission prediction and cost-effective intervention. IEEE J. Biomed. Health Inform 2019 Jul;23(4):1760-1772. [doi: 10.1109/jbhi.2018.2859581]
47. Gönel A. Clinical biochemistry test eliminator providing cost-effectiveness with five algorithms. Acta Clin Belg 2020 Apr 25;75(2):123-127. [doi: 10.1080/17843286.2018.1563324] [Medline: 30585134]
48. Bremer V, Becker D, Kolovos S, Funk B, van Breda W, Hoogendoorn M, et al. Predicting therapy success and costs for personalized treatment recommendations using baseline characteristics: data-driven analysis. J Med Internet Res 2018 Aug 21;20(8):e10275 [FREE Full text] [doi: 10.2196/10275] [Medline: 30131318]
49. Grover D, Bauhoff S, Friedman J. Using supervised learning to select audit targets in performance-based financing in health: An example from Zambia. PLoS One 2019 Jan 29;14(1):e0211262 [FREE Full text] [doi: 10.1371/journal.pone.0211262] [Medline: 30695057]
50. Lee I, Monahan S, Serban N, Griffin PM, Tomar SL. Estimating the cost savings of preventive dental services delivered to Medicaid-enrolled children in six southeastern states. Health Serv Res 2018 Oct 30;53(5):3592-3616 [FREE Full text] [doi: 10.1111/1475-6773.12811] [Medline: 29194610]
51. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 2018 Jun 22;18(1):44 [FREE Full text] [doi: 10.1186/s12911-018-0620-z] [Medline: 29929496]
52. Xie Y, Nguyen QD, Hamzah H, Lim G, Bellemo V, Gunasekeran DV, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. The Lancet Digital Health 2020 May;2(5):e240-e249. [doi: 10.1016/s2589-7500(20)30060-1]
53. Wolf RM, Channa R, Abramoff MD, Lehmann HP. Cost-effectiveness of autonomous point-of-care diabetic retinopathy screening for pediatric patients with diabetes. JAMA Ophthalmol 2020 Oct 01;138(10):1063-1069 [FREE Full text] [doi: 10.1001/jamaophthalmol.2020.3190] [Medline: 32880616]

Abbreviations

AI: artificial intelligence
B2B: business to business
B2C: business to consumers
COGS: cost per good sold
MDR: medical device regulation
R&D: research and development
SAMD: Software as a medical device

Edited by C Lovis; submitted 20.09.21; peer-reviewed by L Le, E van den Akker-van Marle, P Kanzow; comments to author 21.12.21; revised version received 29.03.22; accepted 13.05.22; published 12.08.22.

Please cite as:
Gomez Rossi J, Feldberg B, Krois J, Schwendicke F
Evaluation of the Clinical, Technical, and Financial Aspects of Cost-Effectiveness Analysis of Artificial Intelligence in Medicine: Scoping Review and Framework of Analysis
JMIR Med Inform 2022;10(8):e33703
URL: https://medinform.jmir.org/2022/8/e33703
doi:10.2196/33703
PMID:35969458

Original Paper

Interactive Medical Image Labeling Tool to Construct a Robust Convolutional Neural Network Training Data Set: Development and Validation Study

David Reifs[1], PhD; Ramon Reig-Bolaño[1], PhD; Marta Casals[2], MSN; Sergi Grau-Carrion[1], PhD

[1]Digital Care Research Group, Centre for Health and Social Care, Universitat of Vic-Central University of Catalonia, Vic, Spain
[2]Hospital Santa Creu de Vic, Vic, Spain

Corresponding Author:
David Reifs, PhD
Digital Care Research Group, Centre for Health and Social Care
Universitat of Vic-Central University of Catalonia
Carrer de la Sagrada Família, 7
Vic, 08500
Spain
Phone: 34 938861222
Email: david.reifs@uvic.cat

Abstract

Background: Skin ulcers are an important cause of morbidity and mortality everywhere in the world and occur due to several causes, including diabetes mellitus, peripheral neuropathy, immobility, pressure, arteriosclerosis, infections, and venous insufficiency. Ulcers are lesions that fail to undergo an orderly healing process and produce functional and anatomical integrity in the expected time. In most cases, the methods of analysis used nowadays are rudimentary, which leads to errors and the use of invasive and uncomfortable techniques on patients. There are many studies that use a convolutional neural network to classify the different tissues in a wound. To obtain good results, the network must be trained with a correctly labeled data set by an expert in wound assessment. Typically, it is difficult to label pixel by pixel using a professional photo editor software, as this requires extensive time and effort from a health professional.

Objective: The aim of this paper is to implement a new, fast, and accurate method of labeling wound samples for training a neural network to classify different tissues.

Methods: We developed a support tool and evaluated its accuracy and reliability. We also compared the support tool classification with a digital gold standard (labeling the data with an image editing software).

Results: The obtained comparison between the gold standard and the proposed method was 0.9789 for background, 0.9842 for intact skin, 0.8426 for granulation tissue, 0.9309 for slough, and 0.9871 for necrotic. The obtained speed on average was 2.6, compared to that of an advanced image editing user.

Conclusions: This method increases tagging speed on average compared to an advanced image editing user. This increase is greater with untrained users. The samples obtained with the new system are indistinguishable from the samples made with the gold standard.

(JMIR Med Inform 2022;10(8):e37284) doi:10.2196/37284

KEYWORDS

wound assessment; pressure ulcers; wound tissue classification; labeling; machine learning

Introduction

Skin ulcers are an important cause of morbidity and mortality everywhere in the world [1] and occur due to several causes, including diabetes mellitus, peripheral neuropathy, immobility, pressure, arteriosclerosis, infections, and venous insufficiency. Ulcers are lesions that fail to undergo an orderly healing process and produce functional and anatomical integrity in the expected time (4 weeks to 3 months) [2]. This is usually due to an underlying pathology that prevents or delays healing. Ulcers have a major impact on the patient's life, causing a reduction in the quality of life in physical, emotional [3], and social dimensions. Several contributing and confounding factors are

associated with both the cause and maintenance of ulcers. In addition, care of these wounds requires the expenditure of human and material resources and generates a great economic impact [4]. For these reasons, complex wounds such as ulcers are considered a major global problem.

In most cases, the methods of analysis used nowadays are rudimentary, which leads to errors and the use of invasive and uncomfortable techniques for patients. It is extremely difficult to monitor [5] the evolution of the wound based on the healing process as no data are stored or classified efficiently. Literature covering different algorithms focused on the detection and characterization of wounds is limited and mainly based on the capture of size and depth of the wounds [6,7]. There are many studies that use a convolutional neural network (CNN) to classify the different tissues in a wound [8-11]. However, the process of labeling the images for the training of a CNN in a supervised algorithm is hard work and requires extensive time and effort by a health professional.

In current CNN training models, the labeling of the data set samples is a critical and important phase. In pretrained classification networks, images have been labeled using polygonal contour tools that help detect objects, parts of a body, animals, and so on [12]. For tissue classification, more detailed labeling is required. A wound expert user will have to label the samples, typically using a professional photo editing software. Using the editing tools, this user will paint the different tissues of the wound with predetermined colors (eg, granulated in red, slough in yellow, necrotic in black, and intact skin in blue), pixel by pixel. At the end of the process, 2 files are obtained—1 with the original image and 1 modified with labels drawn with the editing software.

The main goal of this work is to propose an interactive tool for labeling wound samples used for training a CNN to classify different tissues. With this interactive tool, the labeling process is faster, more efficient, and more accurate than with the current manual methods.

Methods

Materials

The collection of the necessary data for labeling was made with a mobile app that uses a standard camera—in our case, a Samsung Galaxy S10 tablet. The data were collected in a health center by health care professionals.

Ethics Approval

The clinical protocol has been approved by the CEIC of the Hospital General de Vic (2019093/PR224).

Proposal

A proposed labeling tool is developed and presented in this study. The results of this application are used for training the CNN model (see the complete working framework in Figure 1). This tool is based on an image editor tool and allows for standard image editing actions such as zoom (Figure 2) and gamma correction (Figure 3). It uses computer vision techniques for tagging and labeling each tissue.

Figure 1. Generic overview of convolutional neural network (CNN) labeling, training, and inference process.

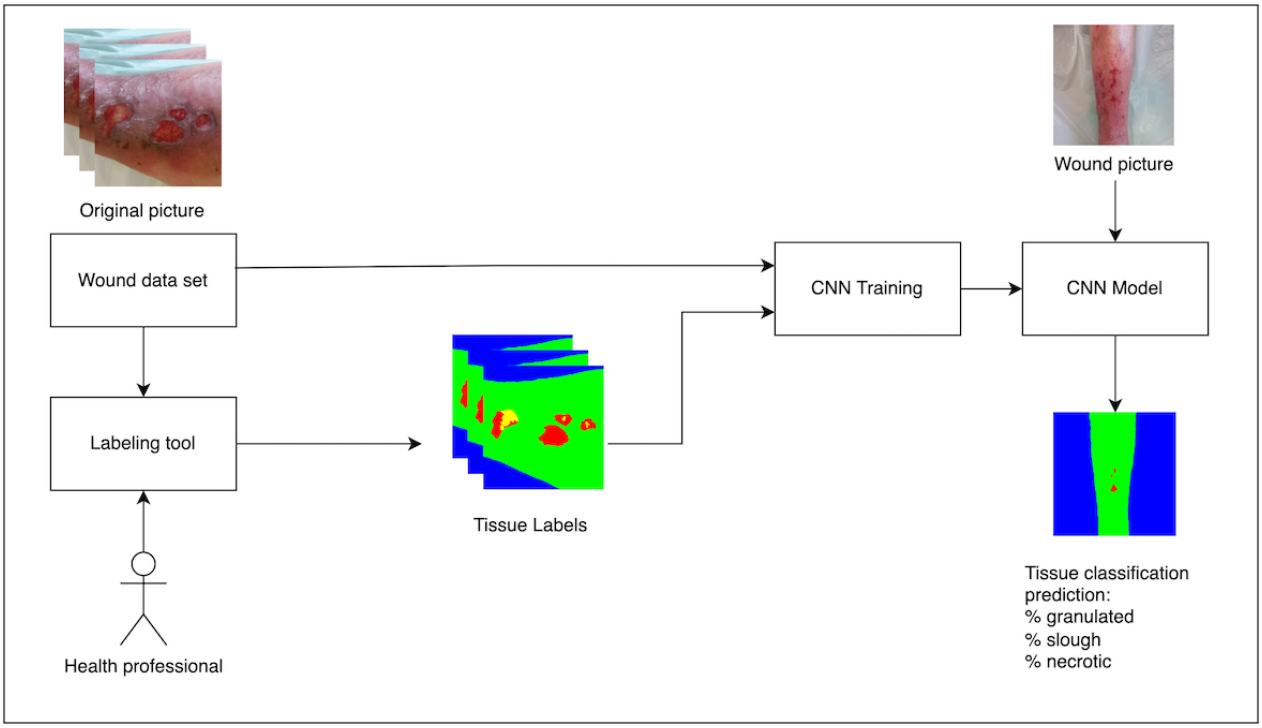


Figure 2. Region selection to apply zoom (left) and the region zoomed (right).

Figure 3. The luminosity of the image can be modified by applying gamma correction. From left to right: original image, gamma value=0.5, and gamma value=2.

The interactive labeling tool can be divided into 2 working stages. In the first stage, the user can choose the part of the image of interest, using the mouse on the original image to define the region of interest (region to label). At the same time, the user can change the image parameters and hyperparameters of the automatic segmentation methods included in the tool.

During the first stage, the tool suggests different partitions of the image the user can select based on which segments best suit the labeling objective and define their class (Figure 4). The partitions are calculated automatically, segmenting the image using computer vision methods and separating the different elements. When the user zooms in on parts of the image to be able to increase the precision in complex areas, the segmentation algorithm recalculates over the zoomed section (Figure 5). The user can also change the hyperparameters (parameters whose value is used to control the algorithm) of the segmentation algorithms to recalculate the partitions and get new proposals (Figure 6).

In the second stage, the user will use the segmentations proposed by the tool to select those that best fit the clinical criteria for tissue classification. The user can make use of sections from different proposals. As the user selects the segmentations, the final labeled image will be drawn in the *Mask* section (Figure 4).

Although the proposed tool allows a desired number of tissues to be tagged, this study was based on the hypothesis of labeling 5 types of tissues: intact skin, slough, necrotic, granulated, and background (or no skin). For this reason, only comparisons between these tissue labels will appear in the results presented.

The segmentation process is based on superpixels and clustering methodologies. It uses different configurations of superpixels and clustering to receive different segmentations of the input image. The resulting segmentations are shown to the user to select the partitions that are closest to the tissue distributions.

In addition, the app has 2 different tools for manual image editing (Figure 7). These tools allow for the correction of mislabeled regions, thus improving the quality of the edges or ambiguous regions hard to segment automatically. The first tool is a brush that allows the user to paint the image using the cursor. The second tool is equivalent to the “magic wand” tool where selecting a pixel in the image causes all the adjacent similar pixels under a threshold to be automatically selected as well.

At the end of the process, the user can obtain a final labeled image where each pixel value is related to the class of the corresponding pixel in the original image (Figure 8).

As mentioned before, the tool uses different computer visual methods based on superpixels (techniques 1, 2, and 3 below) and clustering (technique 4 below). Superpixels are an aggregation of pixels according to similar characteristics between them, such as raw pixel intensity. There are different algorithms and criteria used to measure the similarity between pixels. Clustering is an unsupervised machine learning technique that involves the grouping of data points in a different number of clusters according to the similarity between them.

Figure 4. Main menu view. Left options: brush, wand, back, gamma, quick, Felzenszwalb (FZ), N clusters, and simple linear iterative clustering (Slic). Right options: red (R), yellow (Y), orange (O), black (B), gray (G), blue, move (mv), save, and close.

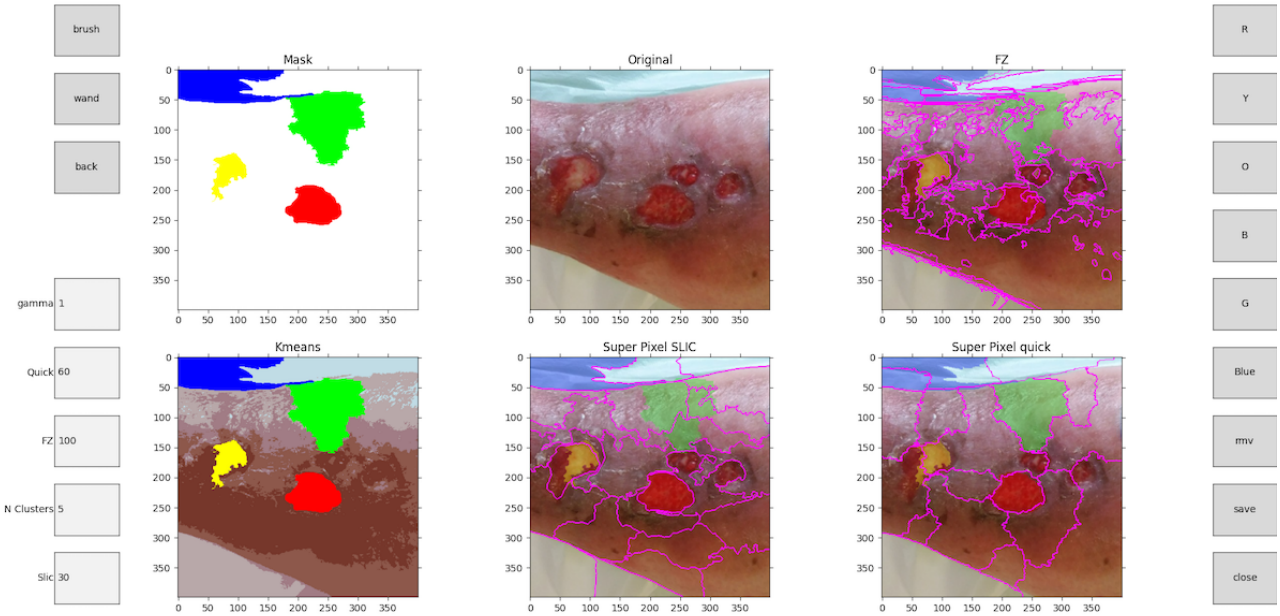


Figure 5. Recalculated partitions from a zoom in the original image. Left options: brush, wand, back, gamma, quick, Felzenszwalb (FZ), N clusters, and simple linear iterative clustering (Slic). Right options: red (R), yellow (Y), orange (O), black (B), gray (G), blue, move (mv), save, and close.

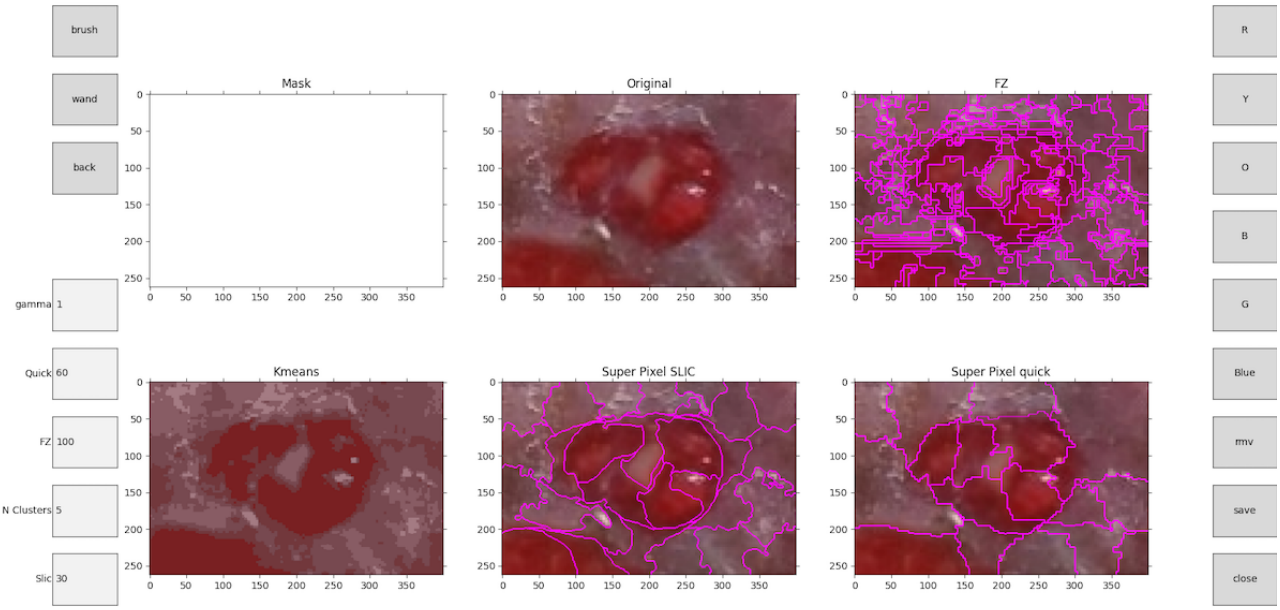


Figure 6. Example of hyperparameters, from left to right: simple linear iterative clustering (SLIC) segmentation with 30 clusters and SLIC segmentation with 100 clusters.

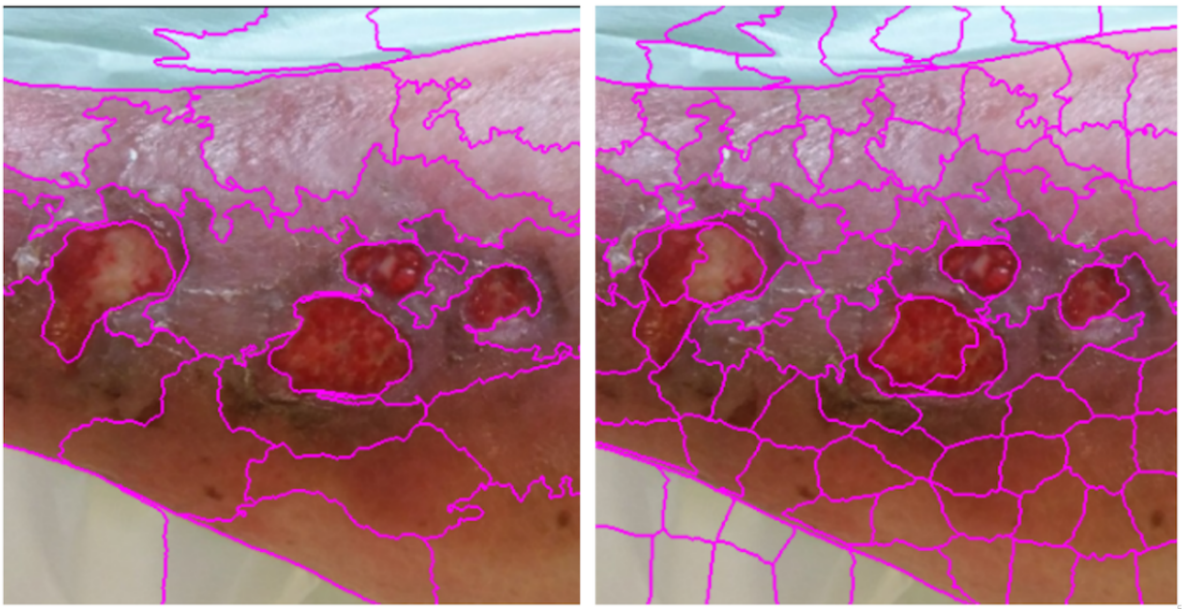


Figure 7. Manual edition tools to classify pixels. RGB: an additive color model with primary colors (red, green, and blue); Std: standard deviation.

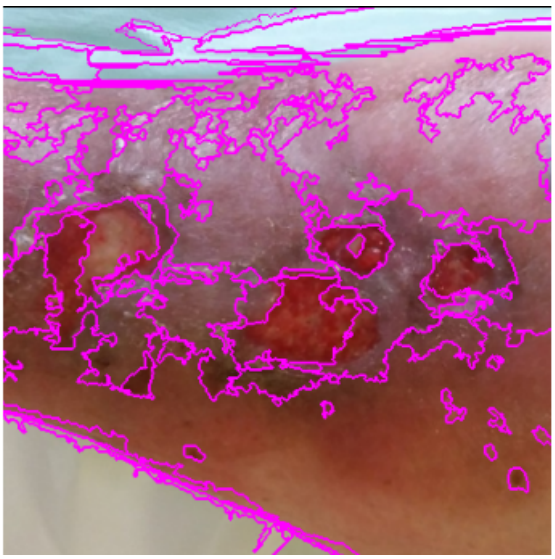
Figure 8. From left to right: original image and labeled image. The classified tissues are intact skin (green), slough (yellow), granulated (red), and background (blue). In this case, there is no presence of necrotic.

Technique 1: Felzenszwalb Efficient Graph-Based Segmentation

Based on superpixels, this technique is a graph-based approach to segmentation [13]. The goal was to develop a computational approach to image segmentation that is broadly useful, much in the way that other low-level techniques such as edge detection are used in a wide range of computer vision tasks. This technique connects elements of the graph according to similarity criteria and a greedy algorithm (Figure 9) to make the boundaries between the different segments more evident.

The similarity criteria used is *Pairwise Region Comparison Predicate*. This predicate is based on measuring the dissimilarity between elements along the boundary of the 2 components. The difference between the 2 components is defined by the minimum weight edge connecting them together.

Figure 9. Felzenszwalb segmentation.

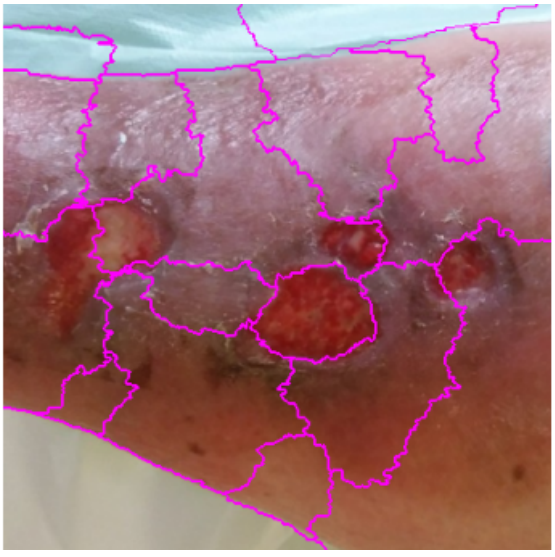


Technique 2: Quickshift Image Segmentation

This technique uses a "Mean-shift" [14] algorithm that segments an RGB (red, green, and blue primary colors) image (or any image with more than one channel) by identifying clusters of pixels in the joint spatial and color dimensions. Segments are local (superpixels) and can be used as a basis for further processing. The cluster approach is carried out over a 5D space defined by the L,a,b values of the CIELAB (International Commission on Illumination) color space and the x,y pixel coordinates (Figure 10).

Mean-shift is a mode-seeking algorithm that generates image segments by recursively moving to the kernel-smoothed centroid for every data point in the pixel feature space, effectively performing a gradient ascent. The generated segments or superpixels can be large or small based on the input kernel parameters, but there is no direct control over the number, size, or compactness of the resulting superpixels.

Figure 10. Quickshift segmentation.



Technique 3: Simple Linear Iterative Clustering Superpixels

This technique's algorithm [15] consists of simple linear iterative clustering, performing a local clustering of pixels in the 5D space defined by the L,a,b values of the CIELAB color space and the x,y pixel coordinates (Figure 11).

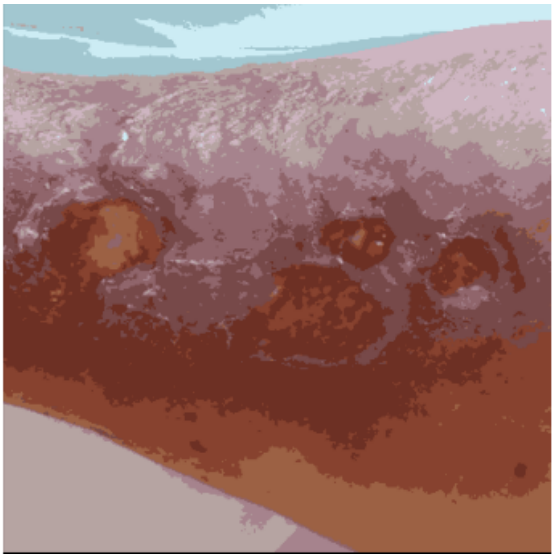
For simple linear iterative clustering, each pixel in the image is associated with the nearest cluster center whose search area overlaps this pixel. After all the pixels are associated with the nearest cluster center, a new center is computed as the average labxy vector of all the pixels belonging to the cluster. We then iteratively repeat the process of associating pixels with the nearest cluster center and recomputing the cluster center until convergence.

Figure 11. Simple linear iterative clustering (SLIC) segmentation.

Technique 4: K-Means Image Segmentation

K-means [16] is a clustering method used to divide a set of data into a specific number of groups. For image segmentation, the clusters are calculated by raw pixel intensities. Image pixels are associated to the nearest centroid using Euclidian distance as a similarity measure (Figure 12).

Figure 12. K-means segmentation.



Results

To evaluate this proposed method, we compared the results obtained by the proposed tool and the results obtained by wound experts using manual segmentation. The manual segmentation was carried out using Gimp, a free cross-platform image editing software, and the experts classified each label pixel by pixel. Specifically, we compared the time used to classify the wound images in each method and the accuracy of our method against the manual one.

Time Evaluation

Table 1 shows the time employed to label each one of the data set samples using the gold standard method versus the proposed method. With the proposed method, the image tagging speed is increased by an average of 2.6 times.

Table 1. Comparison of the time employed to label each sample of the data set with the 2 referred methods, and the speedup achieved with the proposed method; time notation in minutes and seconds (mm:ss).

Sample	Manual method (time)	New method (time)	Speedup achieved
1	10:30	2:47	3.7x
2	05:35	2:30	2.2x
3	07:30	2:06	3.5x
4	09:15	4:11	2.2x
5	06:30	4:42	1.3x
6	13:24	5:38	2.3x
7	03:54	0:41	5.7x
8	03:02	1:16	2.3x
9	02:44	2:09	1.2x
10	07:06	1:29	4.7x
11	04:20	1:30	2.8x
12	04:42	1:25	3.3x
13	03:05	1:01	3.0x
14	06:37	4:02	1.6x
15	03:21	1:15	2.6x
16	02:49	1:38	1.7x
17	03:18	1:35	2.0x
18	05:07	1:48	2.8x
19	03:59	2:50	1.4x
20	03:17	1:14	2.6x

Similarity

Precision, recall, and *F*-score measures are used to evaluate the accuracy of labeling algorithms. The image obtained with the gold standard is taken as ground truth. When tagging an image, it is to be expected that the result obtained will be slightly different each time, even if the same tool and the same criteria are used. It is necessary to be able to evaluate whether the samples labeled with the new method are as similar to the gold standard reference samples as would be other samples made with the same method. Therefore, we relabeled all the gold standard samples to compare the quality of the similarity obtained. The exact correlation between gold standard and new labeling method would be 1.0 (Tables 2 and 3).

Table 2. Comparison between the gold standard and the proposed labeling method.

Tissue	Precision	Recall	*F*-score
No skin (background)	0.9789	0.9824	0.9804
Intact skin	0.9842	0.9867	0.9854
Granular	0.8426	0.9157	0.8753
Base	0.9309	0.8492	0.8838
Necrotic	0.9871	0.7362	0.8387

Table 3. Comparison between the gold standard method samples.

Tissue	Precision	Recall	*F*-score
No skin (background)	0.9919	0.9921	0.9919
Intact skin	0.9938	0.9912	0.9925
Granular	0.8265	0.9377	0.8730
Base	0.9172	0.8821	0.8932
Necrotic	0.9771	0.7622	0.8481

Precision is the relationship between the correctly predicted positive observations and the total expected positive observations. This metric determines how many pixels match out of all the pixels labeled as specific tissue. High precision is related to the low rate of false positives.

Recall, or sensitivity, is the relationship between the correctly predicted positive observations and all positive observations of actual class. This metric determines how many pixels, out of all the pixels that truly matched, were labeled.

F-score provides a single score that balances the concerns of both precision and recall in one value. Therefore, this score considers both false positives and false negatives.

Discussion

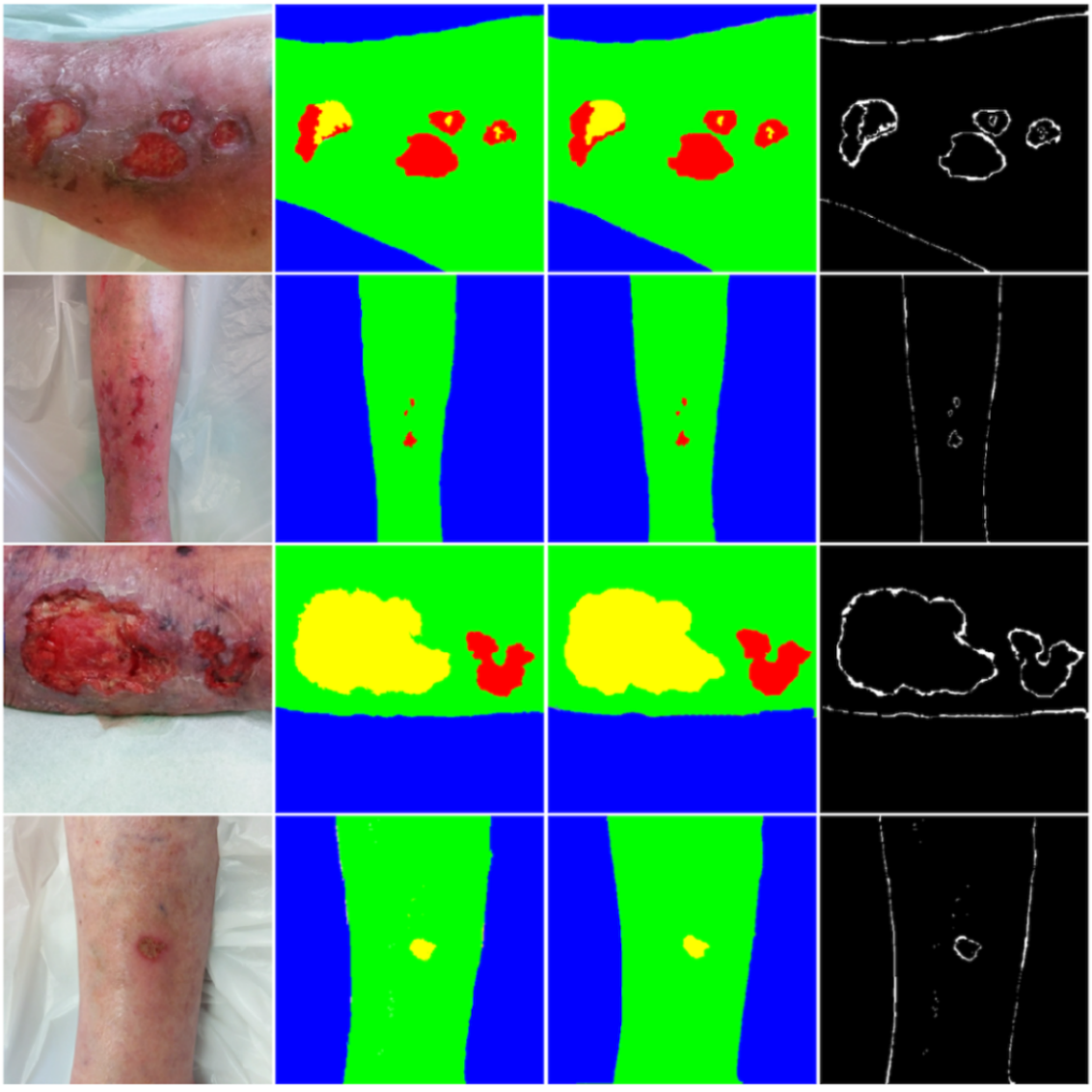
Principal Findings

By analyzing the difference between images labeled with the 2 methods, we see that the discrepancies are found at the edges of the labeling (Figure 13).

This observation is especially relevant for the evaluation of the smallest elements, where the area or perimeter ratio is more significant and can affect the evaluation of similarity. Likewise, any discrepancy of criteria that may exist in the labeling will affect the minority classes to a greater extent. The majority of the classes (no skin and intact skin) have higher *F*-score values than the rest of the classes.

Evaluating the results in Tables 2 and 3, the results obtained with the 2 methods are highly similar, with almost no difference between the comparison of the labels.

Figure 13. From left to right: examples of original image, labeled image with digital method, labeled with gold standard method, and differences between methods.



Conclusions

The proposed method increases tagging speed by an average of 2.6 compared to an advanced image editing user. This gain is larger with untrained users.

The samples obtained with the proposed system are indistinguishable from the samples made with the gold standard.

The incorporation of this type of algorithm will undoubtedly shorten the time required for training a tissue classification network. It provides a tool that can be used by any clinician regardless of their level of knowledge of photo editing. As such, it makes training and using the neural network approach accessible to all in a practical and fast way.

Acknowledgments

We appreciate the collaboration and assistance by the members of the Wound Care department of Hospital de la Santa Creu de Vic who responded to the implementation, assessment, and validation of this new method in their organization. This work has been carried out within the framework of the doctoral program of the University of Vic—Central University of Catalonia.

Conflicts of Interest

None declared.

References

1. Lazarus G, Valle MF, Malas M, Qazi U, Maruthur NM, Doggett D, et al. Chronic venous leg ulcer treatment: future research needs. Wound Repair Regen 2014 Oct 17;22(1):34-42. [doi: 10.1111/wrr.12102] [Medline: 24134795]
2. Coerper S, Beckert S, Küper MA, Jekov M, Königsrainer A. Fifty percent area reduction after 4 weeks of treatment is a reliable indicator for healing--analysis of a single-center cohort of 704 diabetic patients. J Diabetes Complications 2009 Jan;23(1):49-53. [doi: 10.1016/j.jdiacomp.2008.02.001] [Medline: 18394932]
3. Platsidaki E, Kouris A, Christodoulou C. Psychosocial aspects in patients with chronic leg ulcers. Wounds 2017 Oct 03;29(10):306-310 [FREE Full text] [doi: 10.25270/wnds/2017.10.306310] [Medline: 29091039]
4. Nussbaum SR, Carter MJ, Fife CE, DaVanzo J, Haught R, Nusgart M, et al. An economic evaluation of the impact, cost, and Medicare policy implications of chronic nonhealing wounds. Value Health 2018 Jan;21(1):27-32 [FREE Full text] [doi: 10.1016/j.jval.2017.07.007] [Medline: 29304937]
5. Veredas F, Mesa H, Morente L. Binary tissue classification on wound images with neural networks and bayesian classifiers. IEEE Trans Med Imaging 2010 Feb;29(2):410-427. [doi: 10.1109/TMI.2009.2033595] [Medline: 19825516]
6. Restrepo-Medrano JC, Verdú J. Medida de la cicatrización en úlceras por presión: ¿Con qué contamos? Gerokomos 2011 Mar;22(1):1-252 [FREE Full text] [doi: 10.4321/s1134-928x2011000100006]
7. Restrepo-Medrano JC, Verdú Soriano J. Desarrollo de un índice de medida de la evolución hacia la cicatrización de las heridas crónicas. Gerokomos 2011 Dec;22(4):176-183. [doi: 10.4321/s1134-928x2011000400005]
8. Zahia S, Sierra-Sosa D, Garcia-Zapirain B, Elmaghraby A. Tissue classification and segmentation of pressure injuries using convolutional neural networks. Comput Methods Programs Biomed 2018 Jun;159:51-58. [doi: 10.1016/j.cmpb.2018.02.018] [Medline: 29650318]
9. Lucas Y, Niri R, Treuillet S, Douzi H, Castaneda B. Wound size imaging: ready for smart assessment and monitoring. Adv Wound Care (New Rochelle) 2021 Nov 01;10(11):641-661. [doi: 10.1089/wound.2018.0937] [Medline: 32320356]
10. Müller-Linow M, Wilhelm J, Briese C, Wojciechowski T, Schurr U, Fiorani F. Plant Screen Mobile: an open-source mobile device app for plant trait analysis. Plant Methods 2019 Jan 11;15(1):2 [FREE Full text] [doi: 10.1186/s13007-019-0386-z] [Medline: 30651749]
11. Reifs D, Angosto R, Fernandez A, Grau S, Reig-Bolaño R. Tissue segmentation for automatic chronic wound assessment. Front. Artif. Intell. Appl 2019;319:381-384. [doi: 10.3233/FAIA190149]
12. Wei Y, Xia W, Lin M, Huang J, Ni B, Dong J, et al. HCP: a flexible CNN framework for multi-label image classification. IEEE Trans. Pattern Anal. Mach. Intell 2016 Sep 1;38(9):1901-1907. [doi: 10.1109/tpami.2015.2491929]
13. Felzenszwalb PF, Huttenlocher DP. Efficient graph-based image segmentation. International Journal of Computer Vision 2004 Sep;59(2):167-181. [doi: 10.1023/b:visi.0000022288.19776.77]
14. Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Machine Intell 2002;24(5):603-619. [doi: 10.1109/34.1000236]
15. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 2012 Nov;34(11):2274-2282. [doi: 10.1109/TPAMI.2012.120] [Medline: 22641706]
16. Lloyd S. Least squares quantization in PCM. IEEE Trans. Inform. Theory 1982 Mar;28(2):129-137. [doi: 10.1109/tit.1982.1056489]

Abbreviations

CIELAB: International Commission on Illumination
CNN: convolutional neural network

Edited by C Lovis; submitted 14.02.22; peer-reviewed by MS Whiteley; comments to author 24.03.22; revised version received 10.05.22; accepted 31.07.22; published 22.08.22.

Please cite as:
Reifs D, Reig-Bolaño R, Casals M, Grau-Carrion S
Interactive Medical Image Labeling Tool to Construct a Robust Convolutional Neural Network Training Data Set: Development and Validation Study
JMIR Med Inform 2022;10(8):e37284
URL: https://medinform.jmir.org/2022/8/e37284
doi:10.2196/37284
PMID:35994311

Original Paper

An Efficient Method for Deidentifying Protected Health Information in Chinese Electronic Health Records: Algorithm Development and Validation

Peng Wang[1*], MEng; Yong Li[2*], MEng; Liang Yang[3], MEng; Simin Li[3], MEng; Linfeng Li[3], PhD; Zehan Zhao[4], MEng; Shaopei Long[2], BEng; Fei Wang[5], MEng; Hongqian Wang[5], MEng; Ying Li[5], MSc; Chengliang Wang[1], PhD

[1]College of Computer Science, Chongqing University, Chongqing, China

[2]School of Computer Science, South China Normal University, Guangzhou, China

[3]Yidu Cloud Technology Inc, Beijing, China

[4]School of Software & Microelectronics, Peking University, Beijing, China

[5]Medical Big Data Center of Southwest Hospital, Chongqing, China

[*]these authors contributed equally

Corresponding Author:
Chengliang Wang, PhD
College of Computer Science
Chongqing University
No. 55, South University Town Road
Shapingba District
Chongqing, 400030
China
Phone: 86 18983055830
Email: wangcl@cqu.edu.cn

Abstract

Background: With the popularization of electronic health records in China, the utilization of digitalized data has great potential for the development of real-world medical research. However, the data usually contains a great deal of protected health information and the direct usage of this data may cause privacy issues. The task of deidentifying protected health information in electronic health records can be regarded as a named entity recognition problem. Existing rule-based, machine learning–based, or deep learning–based methods have been proposed to solve this problem. However, these methods still face the difficulties of insufficient Chinese electronic health record data and the complex features of the Chinese language.

Objective: This paper proposes a method to overcome the difficulties of overfitting and a lack of training data for deep neural networks to enable Chinese protected health information deidentification.

Methods: We propose a new model that merges TinyBERT (bidirectional encoder representations from transformers) as a text feature extraction module and the conditional random field method as a prediction module for deidentifying protected health information in Chinese medical electronic health records. In addition, a hybrid data augmentation method that integrates a sentence generation strategy and a mention-replacement strategy is proposed for overcoming insufficient Chinese electronic health records.

Results: We compare our method with 5 baseline methods that utilize different BERT models as their feature extraction modules. Experimental results on the Chinese electronic health records that we collected demonstrate that our method had better performance (microprecision: 98.7%, microrecall: 99.13%, and micro-F1 score: 98.91%) and higher efficiency (40% faster) than all the BERT-based baseline methods.

Conclusions: Compared to baseline methods, the efficiency advantage of TinyBERT on our proposed augmented data set was kept while the performance improved for the task of Chinese protected health information deidentification.

(JMIR Med Inform 2022;10(8):e38154) doi:10.2196/38154

KEYWORDS

EHR; PHI; personal information; protected data; protected information; patient information; health information; de-identification; de-identify; privacy; TinyBert; model; development; algorithm; machine learning; CRF; data augmentation; health record; medical record

Introduction

Background

With the boost in information technology, electronic health records (EHRs) have been widely adopted and applied in many hospitals and medical institutes. The vast advantages of EHRs include easy storage and management, and they can greatly increase the speed of information retrieval. They can provide abundant clinical and medical information on various diseases, and this information can potentially provide clinicians with evidence for decision-making. However, the private information of many individuals is stored in the EHRs. The incorrect usage of EHRs may cause privacy leakage, leading to serious problems. In order to standardize the use of EHRs and protect individual privacy, many projects, such as the i2b2 challenge, in 2014 [1], and the CEGS N-GRID challenge, in 2016 [2], have been launched. An intuitive method to prevent privacy leakage is deidentifying the protected health information (PHI) [3] in EHRs before information processing. PHI is classified into 18 different types by the US Health Insurance Portability and Accountability Act [4], such as name, ID number, location, date, and age. The process of deidentifying PHI can be divided into 2 steps: locating the PHI in the EHR and replacing it with information that is not sensitive. Accordingly, the deidentification procedure can be treated as a named entity recognition (NER) task [5].

Related Work

In the past few decades, rule-based [6,7] and machine learning–based [3,8,9] approaches have been the mainstream approaches to identifying entities in sentences or documents. Rule-based methods utilize special semantic dictionaries to establish a set of regular expressions [4,5] to extract PHI from EHRs. However, these methods are labor intensive and time consuming, with poor generalization capability. Machine learning methods based on the principles of statistics could automatically detect PHI in EHRs by utilizing manually extracted text features [3,10]. For example, Jian et al [11] designed a set of regular expressions based on the characteristics of Chinese EHRs to filter sentences with sparse PHI, then used the filtered sentences to train a conditional random field (CRF) model for PHI recognition. Du et al [12] manually extracted lexical and dictionary features of PHI from Chinese EHRs to train a CRF model and utilized regular expressions to capture missed features using the CRF. On the basis of the extracted lexical features, Zhang et al [13] employed a long short-term memory (LSTM) method to learn the features of PHI sentences. However, these machine learning–based methods heavily depend on high-quality manual selection of features, which requires a great amount of domain expertise. In recent years, many deep learning models have been applied to the deidentification of PHI. Compared to rule-based and machine learning–based methods, deep learning models could extract features automatically from input words or text vectors [14,15]. However, deep learning–based models require very large annotated data sets for model training to avoid overfitting. To solve this problem, it is tempting to perform data augmentation [16,17] when facing data set insufficiency.

Currently, deidentifying PHI with deep neural networks remains a greater challenge for Chinese-language clinical texts than for other languages [18]. At present, much existing research on PHI deidentification has been done on the English-language corpus. Increasing performance has been achieved for rule-based, machine learning–based, deep learning–based, and hybrid approaches [19,20]. However, the direct application of these methods to Chinese clinical texts for PHI deidentification may result in unsatisfactory results. The huge differences in morphological features between Chinese and English make it futile to construct rules and dictionaries. For example, there is no delimiter in the middle of a sentence in Chinese, because the basic morpheme that expresses meaning in Chinese consists of more than one word. Additionally, Chinese grammar is more flexible, and some words can exist as multiple parts of speech. In addition, the absence of capitalization makes it difficult to locate personal names in Chinese through specific rules. As a result, deep neural networks require a very large Chinese biomedical corpus for learning the high level contextual semantic features of Chinese. However, annotating a large amount of Chinese data for network training is costly, labor intensive, and time consuming. Thus, there is a great need for the ability to train deep neural networks on limited-size annotated Chinese data sets. To reduce model dependence on limited training data, an intuitive method would be to fine-tune a model that has been pretrained with a Chinese corpus with the target-specific downstream data set. However, there are two limitations on applying pretrained language models to downstream tasks. First, if the pretraining tasks and the target tasks are not domain matched, the pretraining model may impair the performance of the target tasks [21]. Second, there can be overfitting issues when there is not enough data for fine tuning the downstream tasks.

Objective

In this paper, we propose a deep neural network that uses TinyBERT [22] and a CRF model for Chinese PHI deidentification. TinyBERT as used in our model is distilled from a BERT (bidirectional encoder representations from transformers)-based model that was pretrained on a Chinese corpus. It has two advantages: it can overcome the differences in the morphological features of Chinese and English, and it has fewer parameters, which should prevent the deep learning model from overfitting when training on small-scale Chinese EHR data sets. In addition, we propose a hybrid data-augmentation method that uses data augmentation with a generation approach (DAGA) [23] and mention replacement (MR) [24] to create more training data. The enhanced data set

assists the neural network in overcoming overfitting and enhances the generalizability of the deep neural networks.

Methods

The PHI Recognition Model

In this paper, a new model that integrates TinyBERT [22] and a CRF model [25] is proposed for PHI recognition in Chinese EHRs. As shown in Figure 1, this model utilizes TinyBERT as the feature extraction module and the CRF model as the prediction module. The words in the sentences of an EHR are first tokenized, and the lengths of the sentences are fixed to 128. They are then input to the embedding module of TinyBERT to generate word embeddings, position embeddings, and token-type embeddings. The 3 embedding matrices are added together as input to the feature encoder, consisting of cascaded self-attention blocks for text feature extraction. With the self-attention mechanism, the model captures long-distance interdependent features in sentences and learns the semantics of the sentences. The feature extraction module outputs a series of probabilities for sequence labels, which are regarded as the emission scores of the CRF model. After that, the text features are input to the CRF module for label prediction.

TinyBERT is a light structure which is generated with the transformer-layer distillation method from the base BERT [26]. The structures to be distilled are an embedding layer, multiple transformer layers, and a prediction layer. The details of the model distillation process are shown in Figure 2. Assuming that the base BERT is the teacher module and has M transformer layers, TinyBERT is the student module and has N transformer layers, where M = k × N. In the distillation process, the model learns knowledge through a knowledge distillation (KD) function between the indices from the teacher module to student module, as shown in equation 1:

$$\theta_S(n) = g(k,\theta_T(m)) \quad \textbf{(1)}$$

where $\theta_S(n)$ denotes the parameters of the student module with n transformer layers, $\theta_T(m)$ denotes the parameters of the teacher module with m transformer layers, and g(•) denotes the knowledge mapping function from the teacher module to the student module. Formally, g(•) is optimized through minimizing the distillation loss (L(distillation)), which is summed by the transformer layer loss (L(tr)), the embedding layer loss (L(emb)), and the prediction loss (L(pr)). To generate the TinyBERT model, training sequences with a length of 1 were simultaneously input to the teacher module and the student module for label prediction, and the distillation loss was then minimized in the training process, which can be calculated from equation 2 to equation 5, as follows:

$$L(emb) = ||E^S,E^TW_e||_2 \quad \textbf{(3)}$$

$$L(pr) = \text{cross\_entropy}\ (Z^T,Z^S) \quad \textbf{(4)}$$

$$L(distillation) = L(tr) + L(emb) + L(pr) \quad \textbf{(5)}$$

where h is the number of attention heads.  denote the *i-th*-layer attention map values, the output feature maps of the transformer blocks, the output of the embedding layers, and the predicted logic vectors of the student module, respectively.  denote the *i-th*-layer attention map values, the output feature maps of the transformer blocks, the output of the embedding layer, and the predicted logic vectors of the teacher module, respectively. W_h and W_e denote the linear transformation matrices, and  , where $\in \{A, H,E,Z\}$.

After the knowledge distillation process, the parameters of the obtained TinyBERT were dramatically shrunk, while reserving most of the knowledge of the base BERT. Our model utilizes the text features output by the last TinyBERT encoder to finally obtain the predicted labels through a classifier, such as a softmax function. However, the softmax function regards each vector as independent and ignores correlations between word labels in a sentence; thus, some unreasonable results may be predicted. To eliminate this issue, we introduced the CRF model to build the dependencies and constraints within annotated sequences. Instead of assuming that the current label of a token depends only on the current label or the current label depends only on a previous label, the CRF model breaks the limitations of local token dependencies and focuses on the whole sentence. Specific dependency rules that can be learned in the NER task are shown in Figure 2.

The label for the first word in a sentence should start with “B-” or “O,” not “I-.” In the mode that “B - label_1 I - label_2 I - label_3 I -...” there should be the same named entity tag for label_1, label_2, and label_3. Based on this rule, it is easy to exclude wrong predictions, such as “B-Person I-Organization...” Based on the observations, the CRF model can define an equation to score a predicted sequence label of the input sentence according to the dependency rules in equations (6) to (8):

$$\text{score}(X|s) = \text{emission\_score} + \text{transition\_score} \quad \textbf{(6)}$$

where s denotes the input sentence, $_{i,label}$ denotes the score of the predicted labels of the *i-th* word in the sentence s, and $_{labeli \rightarrow labelj}$ denotes the score of transferring the $label_i$ to $label_j$ of the word , respectively. In our method, the *emission_score* is obtained from the output of the TinyBERT module, and the *transition_score* is calculated by the CRF module with the contextual information in the sentence. To maximize the probability of correct predicted sequence labels, the exponent and standardization among all the predicted scores are calculated according to equation 9:

Therefore, the loss function for optimizing our model can be defined as equation 10:

Figure 1. The proposed model for deidentifying protected health information in Chinese electronic health records. BERT: bidirectional encoder representations from transformers; CRF: conditional random field; FFN: feed-forward network; MHA: multi-head attention; PER: personal name.

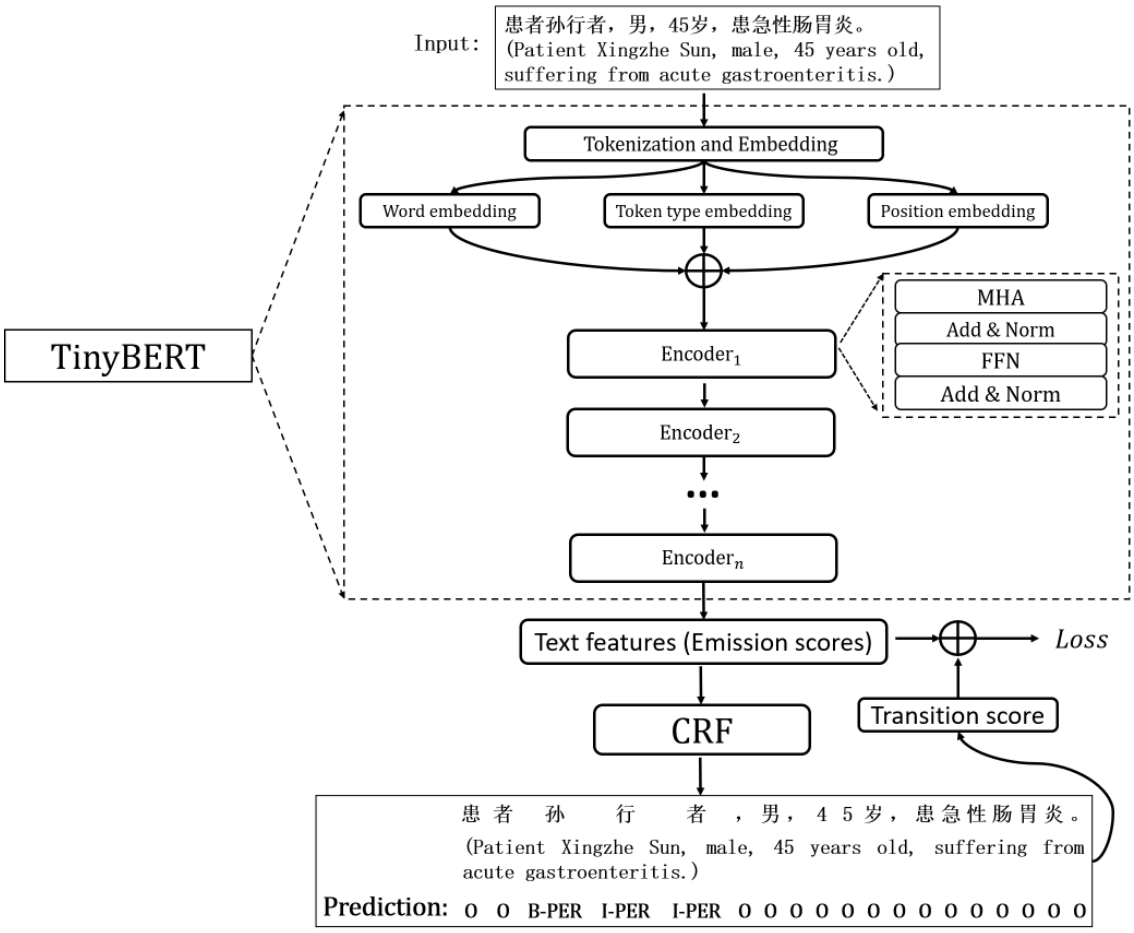
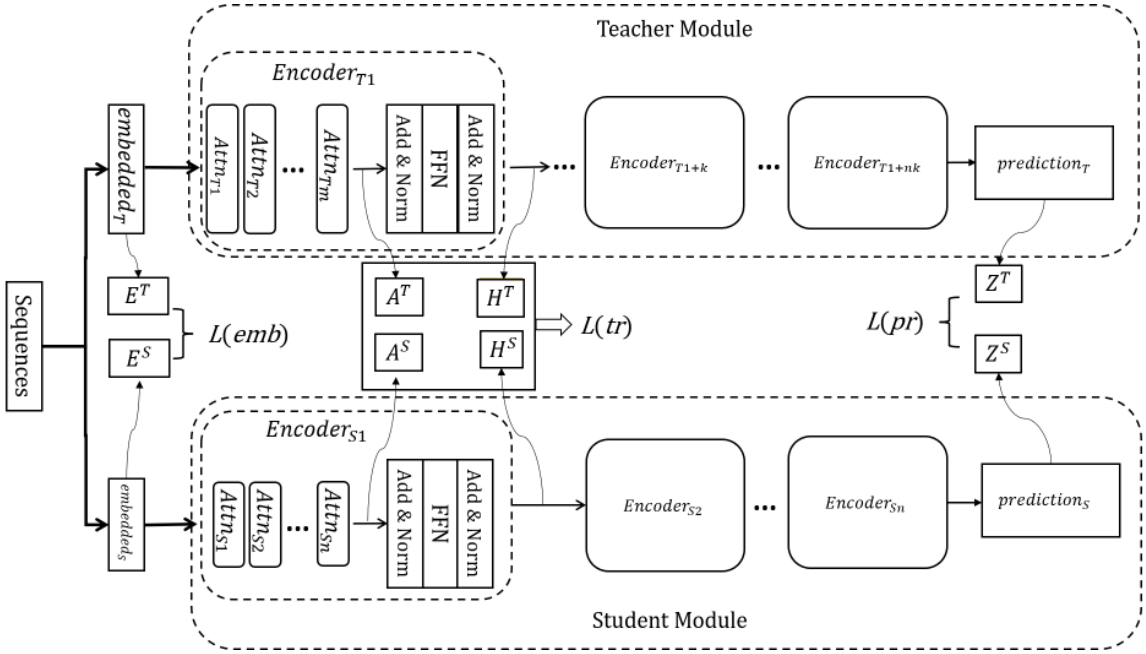


Figure 2. The TinyBERT knowledge distillation process used in our model. FFN: feed-forward network. Attn: attention layer; L(emb): embedding loss; L(tr): transformer layer loss; L(pr): prediction loss; A: attention map values; Z: predicted logic vectors; S: student network; T: teacher network.



A Hybrid Data Augmentation Method

Formally, there is a trade-off between the performance and efficiency of a deep neural network. The performance of a network degrades, while its efficiency is enhanced, when the parameters are compressed. In practice, a network compresses the number of transformer layers and word embedding dimensions to improve efficiency, but this also results in the

ability of feature extraction becoming inferior. To keep its efficiency without degrading its performance, an intuitive method is to fine-tune it on a large data set. Unfortunately, the generation of a sufficient, high-quality data set is challenging. As discussed in previous reports [23,24], augmenting data with noise may enhance the robustness of the models on tasks at the sentence level, such as text classification and emotional judgment, but it harms the performance of tasks at the token level, such as NER. This situation indicates that augmented data should contain as little noise as possible. Furthermore, the research of Dai et al [27] indicates that hybrid data augmentation outperforms any single method of data augmentation, on average. Inspired by this work, we propose a new hybrid data augmentation method, which combines DAGA [23] and MR [24] to enhance original data for task-specific fine-tuning. The DAGA is used to increase the size of the training set so as to avoid overfitting, while MR is used to enable a network to learn different representations of entities.

Unlike other data augmentation methods, a DAGA generates new synthetic data from scratch without relying on WordNet (Princeton University) or other external dictionaries, which could make it more useful for limited-resource languages. It mixes entity labels and word tokens together to create a linear sentence. An example is shown in Figure 3. The generated linear sentences are input to a word generation network (such as an LSTM or BERT) to learn the distribution of words and tags. Given a sequence of tokens (w_1,w_2,...w_t,...,w_N) to the networks, where N denotes the length of the sequence, the networks learn the hidden states of each word in this sequence with equation 11:

$$h_t = Me_t \quad \textbf{(11)}$$

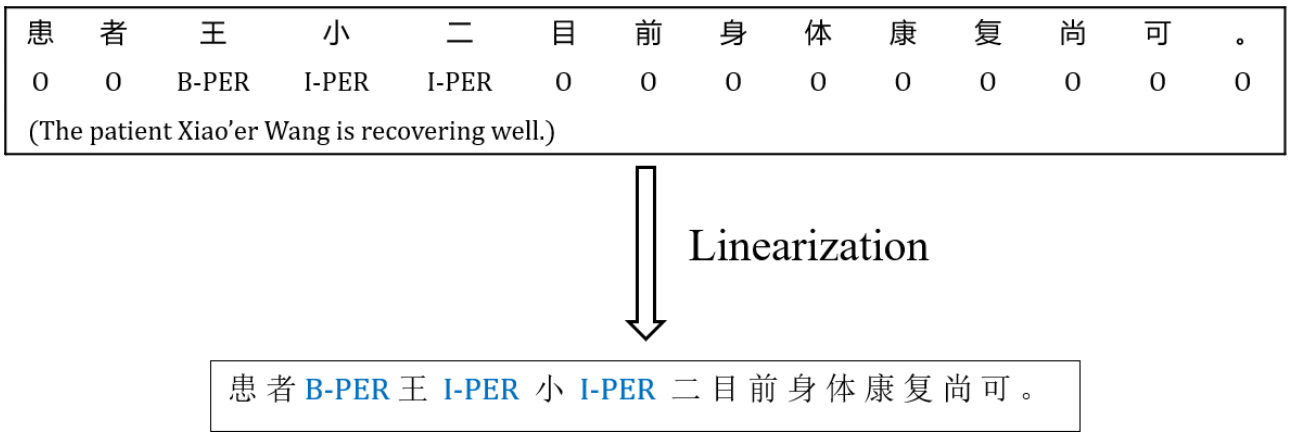
where M denotes the learnable weight matrix in the word-generation networks and e_t denotes the embedding matrix of the input words. The word-generation networks learn to predict the tags of the next token in the sequence by maximizing the probability calculated by equation 12 in the process of training:

where *V* denotes the size of the vocabulary, *i** denotes the index of the word w_t in the vocabulary, and h_{t-1},*i** denotes the *i-th* element of h_{t-1}. In this way, the objective function for obtaining the parameter θ is described in equation 13:

The paired token-label linear sentences promote learning by the networks of the context relationship between parts of speech, so the distribution of the generated synthetic data is closer to the original data, thereby introducing less noise during data augmentation. In addition, the generated synthetic data introduces more diversity to enhance the robustness of the model.

However, our originally collected data set may contain sentences that have fewer entities and more "O" tag words. According to equation 13, a DAGA heavily relies on contextual semantic information for sentence generation. Hence, only applying a DAGA to the originally collected data set for data augmentation may cause an entity sparsity issue, which is not conducive to a model for learning rich data features. To mitigate this, we introduced MR as another supplementary data augmentation method. For each labeled entity in a sentence, we formulated a binomial distribution to determine if the entity should be replaced. The formula outputs a probability *P*, and the entity is chosen for replacement by another entity from the training set that has the same entity type when *P*>.5. Otherwise, the entity remains in the original sentence. However, due to the small size of the originally collected data set, applying only MR for data augmentation easily generates duplicate data, which may cause oversampling in the training process, resulting in overfitting of the model. Therefore, we merged a DAGA and MR together to augment the data set.

Figure 3. An example of the data augmentation with a generation approach linearization operation in our data augmentation method. PER: personal name.



Results

Data

The raw EHRs contain patient history, current illness, an admission summary, a daily record of the disease course, the diagnosis, treatment processes, and a discharge summary. The EHRs were all collected from local hospitals in Chongqing City, China. In this paper, we aim to identify protected information in the EHRs, such as the organization (ORG), location (LOC), dates (DAT), and personal names (PER), including the names of patients and doctors.

Manually annotating the raw data is a time-consuming and labor-intensive task, and the data are usually insufficient for disease-specific research, especially for rare diseases. Inspired by past research [28,29], we utilized a deep learning method for the raw data annotation. In this method, all the raw data are randomly split into 2 parts. The first part is called the "mini data set" (containing about 10% of all the raw data) and the other part is called the "formal data set." We invited 2 professional clinicians to annotate all the PHI manually in the mini data set. Then, we fed the annotated mini data set to the base BERT with a CRF model to fine-tune it. Next, we switched the base BERT with the CRF model from a training mode to a test mode to predict the PHI in the formal data set. However, there may have been some incorrect predictions (also called bad cases) in the formal data set. Thus, we manually reviewed the predicted PHI in the formal data set and corrected the bad cases. In the end, we obtained a complete annotated data set with PHI labels. After that, private information, such as patient names, was replaced with random surrogates.

Experiment Settings

We randomly split the raw annotated data set into a training set (denoted as $data_{raw}$), an evaluation set, and a test set at a ratio of 6:2:2. Statistically, there were a total of 2707, 1424, 509, and 5046 labeled PER, ORG, LOC, and DAT entities, respectively. Our data augmentation method was applied to $data_{raw}$ to create a new training set named the "hybrid augmented data set," denoted as $data_{DAGA+MR}$. For comparison, we separately applied a DAGA and MR to the $data_{raw}$ to create 2 additional training sets, denoted as $data_{DAGA}$ and $data_{MR}$. The evaluation set was used for verifying performance in the training process and the test set was used for testing the performance of our proposed model and other baseline methods. Detailed statistical information on our hybrid augmented data set and the raw data set for each type of entity are shown in Table 1.

We retained the CRF module and replaced the feature extraction module of our model with other modules. These modules included 2 recurrent neural network (RNN)-based models, including BiLSTM [30], gated recurrent units (GRU) [31], and 7 BERT-based models, including base BERT [26], Chinese-BERT-wwm [32], Chinese-BERT-wwm-ext [32], Chinese-BERT-base [33], and Chinese-BERT-large [33], and as baselines, PCL-BERT [34] and PCL-BERT-wwm [34]. Detailed settings for each benchmark model are listed in Table 2. For the evaluation metrics, we used precision, recall, and the F1 score to evaluate the overall performance in the data sets, calculated according to equations (14) to (16), as follows:

where TP, FP, and FN denote true positive number, false positive number, and false negative number, respectively. In practice, the experiments with the base BERT, Chinese-BERT-wwm, Chinese-BERT-wwm-ext, Chinese-BERT-base, Chinese-BERT-large, and TinyBERT models were conducted on a computer with an Intel Xeon central processing unit (CPU) (E5-2620, v3, 2.40 GHz) with 128 GB memory. The experiments with the GRU, BiLSTM, PCL-MedBERT, and PCL-MedBERT-wwm were conducted on an Nvidia RTX3090 graphics processing unit (GPU).

Table 1. Statistical information for the raw data and hybrid augmented data for each type of entity.

Entity types	Training set, n				Evaluation set (original), n	Test set (original), n
	Original	DAGA[a]	MR[b]	Total		
PER[c]	1448	4327	2892	8667	631	628
LOC[d]	302	1384	589	2275	102	105
ORG[e]	846	2188	1692	4726	275	303
DAT[f]	3013	7412	6011	16,436	999	1034
Total	5609	15,311	11,184	32,104	2007	2070

[a]DAGA: data augmentation with a generation approach.

[b]MR: mention replacement.

[c]PER: personal name.

[d]LOC: location.

[e]ORG: organization name.

[f]DAT: date.

Table 2. Settings for each benchmark.

Models	Settings	Parameters, n	Description
Gated recurrent units	1 layer,[a] 512 dims[b]	2,190,000	The parameters were randomly initialized.
BiLSTM[c]	1 layer, 512 dims	2,210,000	The parameters were randomly initialized.
Base BERT[d]	12 layers, 768 dims, 12 heads[e]	110,000,000	The base BERT was pretrained on the English Wikipedia corpus.
Chinese-BERT-wwm	12 layers, 768 dims, 12 heads	110,000,000	The base BERT was pretrained on the Chinese Wikipedia corpus with a whole word masking training strategy.
Chinese-BERT-wwm-ext	12 layers, 768 dims, 12 heads	110,000,000	The base BERT was pretrained on the Chinese Wikipedia corpus, news, and question-answer pairs with a whole word masking training strategy.
Chinese-BERT-base	12 layers, 768 dims, 12 heads	147,000,000	The base BERT was pretrained on the Chinese Wikipedia corpus with char, glyph, and pinyin embedding.
Chinese-BERT-large	24 layers, 1024 dims, 12 heads	374,000,000	The base-BERT-large model with more layers and larger dims was pretrained on the Chinese Wikipedia corpus using char, glyph, and pinyin embedding.
PCL-MedBERT	12 layers, 768 dims, 12 heads	110,000,000	A BERT model that was pretrained on the Chinese medicine corpus.
PCL-MedBERT-wwm	12 layers, 768 dims, 12 heads	110,000,000	A BERT model that was pretrained on the Chinese medicine corpus with whole word masking training.
TinyBERT	6 layers, 768 dims, 12 heads	67,000,000	A BERT distilled from the Chinese-BERT-wwm.

[a]Layer: transformer blocks.

[b]Dims: embedding dimensions.

[c]LSTM: long short-term memory.

[d]BERT: bidirectional encoder representations from transformers.

[e]Heads: attention heads.

Experiment Results

The performance of our model compared with the baseline models on the test set is reported in Table 3. After fine-tuning $data_{raw}$, base BERT obtained the best precision (98.55%), while PCL-MedBERT-wwm achieved the best recall (99.18%) and F1 score (98.8%). However, after fine-tuning the models on the hybrid augmented data set, our model obtained the best scores for precision (98.7%), recall (99.13%), and F1 score (98.91%), representing increases of 0.86% for precision, 0.53% for recall, and 0.69% for F1 score compared with $data_{raw}$. Nevertheless, the other baseline models gained improved performance after fine-tuning on the hybrid augmented data set compared to $data_{raw}$. Furthermore, the overall performance of the 2 RNN-based models was inferior to most of the BERT-based models, and the BiLSTM outperformed the GRU on precision, recall, and F1 score by 2.2%, 2.95%, and 2.58%, respectively, after training on $data_{raw}$, and by 1.63%, 2.37%, and 2%, respectively, after training on the hybrid augmented data set.

It is worth noting that the performance of Chinese-BERT-base and Chinese-BERT-large were worse than the other BERT-based benchmark models after fine-tuning on $data_{raw}$. The improvement of these 2 models surpassed the other models after fine-tuning on the augmented data set. Compared to fine-tuning on $data_{raw}$, Chinese-BERT-base achieved increases of 13.94% for precision, 11.69% for recall, and 12.84% for F1 score, and Chinese-BERT-large achieved increases of 1.85% for precision, 0.87% for recall, and 1.36% for F1 score.

In order to further evaluate the effectiveness of our hybrid data augmentation method, we conducted an ablation study through fine-tuning each benchmark on $data_{DAGA}$ and $data_{MR}$. The results are shown in Table 4. Each metric of our model fine-tuned on either $data_{DAGA}$ or $data_{MR}$ performed better than when fine-tuned on $data_{raw}$. The precision, recall, and F1 score improved 0.48%, 0.43%, and 0.46%, respectively, after fine-tuning our model on $data_{MR}$, and improved 0.34%, 0.48%, and 0.38%, respectively, after fine-tuning on $data_{DAGA}$. However, fine-tuning on a single augmented data set could not ensure that our model outperformed other baseline methods on each metric. Overall, the PCL-MedBERT-wwm obtained the best precision and F1 score after fine-tuning on $data_{MR}$ and $data_{DAGA}$.

It is worth noting that the results of some baseline benchmarks degraded after fine-tuning on $data_{MR}$ or $data_{DAGA}$. For example, after fine-tuning the models on $data_{MR}$, the performance of PCL-MedBERT decreased 0.19% for precision, recall, and F1 score, and the performance of base BERT decreased 0.3%, 0.1%, and 0.2% for precision, recall, and F1 score, respectively. The situation was similar for Chinese-BERT-wwm-ext and Chinese-BERT-large. The performance of Chinese-BERT-wwm-ext decreased 0.29% for precision and 0.05% for F1 score, and the performance of Chinese-BERT-large decreased 0.47% for precision.

Nevertheless, the performance of all the benchmark models improved after fine-tuning on our hybrid augmented data set, which proves the effectiveness of the proposed hybrid augmentation method.

We compared the performance on various entity types of our model after fine-tuning it on different data sets. As shown in Table 5, fine-tuning our model on either a single augmented data set or the hybrid augmented data set improved the performance for each entity type, which demonstrates the effectiveness of our proposed data augmentation strategy. It is worth noting that our model could not achieve the best performance for the PER and DAT entity types after fine-tuning on the hybrid augmented data set. For the DAT type, the best results were obtained after fine-tuning our model on $data_{MR}$, with increases of 0.1% for precision, 0.29% for recall, and 0.19% for F1 score compared to the hybrid augmented data set. For the PER type, the best precision was obtained after fine-tuning our model on $data_{DAGA}$; this was 0.16% higher than for $data_{DAGA+MR}$.

To investigate the effect of data volume on our proposed model, we built 4 additional training sets with different data volume, denoted as  ,  ,  , and  . These symbols and their corresponding meanings are listed in Table 6.

The results of our model after fine-tuning on the 4 additional training sets are shown in Table 7. From the table, we can observe that our model fine-tuned on  only obtained performance of 91.33%, 95.26%, and 93.26% for precision, recall, and F1 score, respectively. When the volume of raw data increased to 50%, the performance improved greatly. Furthermore, the performance of our model fine-tuned on either  or  was better than when fine-tuned on $data_{Raw}$,  , or . Moreover, our model obtained better performance after fine-tuning on  than on  . The results also indicate that the less raw data we had, the more the performance of our model improved after fine-tuning on the hybrid augmented data set.

The time used by the different devices for all models that used the test set (including 1500 samples) was recorded for an efficiency evaluation. All the benchmarks ran a forwarded process on the test set; the results are shown in Table 8. Our model achieved the highest efficiency among all the BERT-based benchmarks: 158.22 seconds of CPU time and 62.39 seconds of GPU time. From the table, we can observe that the efficiency increase for CPU time was greater than for GPU time. The more limited were the computing resources, the greater was the efficiency improvement. These results show that our proposed method had higher efficiency with higher performance. Although the efficiency of the GRU and LSTM models was better than our model, the performance of these models for precision, recall, and F1 score was worse.

Table 3. Comparison of each benchmark model after fine-tuning on the raw data and the hybrid augmented data. Italics indicate the best performance.

Models	$Data_{raw}$			$Data_{DAGA+MR}$[a]		
	P,[b] %	R,[c] %	F1,[d] %	P, %	R, %	F1, %
Gated recurrent units	94.92	93.04	93.97	95.9	95.02	95.46
BiLSTM[e]	97.12	95.99	96.55	97.53	97.39	97.46
Base BERT[f]	*98.55*	98.7	98.63	98.65	98.85	98.75
Chinese-BERT-wwm	98.35	98.5	98.43	98.5	98.90	98.7
Chinese-BERT-wwm-ext	98.4	98.5	98.45	98.65	98.90	98.78
Chinese-BERT-base	82.92	85.36	84.12	96.86	97.05	96.96
Chinese-BERT-large	95.42	95.7	95.56	97.27	96.57	96.92
PCL-MedBERT	98.37	99.08	98.72	98.36	98.79	98.58
PCL-MedBERT-wwm	98.42	*99.18*	*98.8*	98.46	98.89	98.67
Our model	97.84	98.6	98.22	*98.7*	*99.13*	*98.91*

[a]DAGA+MR: data augmentation with a generation approach and mention replacement.

[b]P: precision.

[c]R: recall.

[d]F1: F1 score.

[e]BiLSTM: bidirectional long short-term memory.

[f]BERT: bidirectional encoder representations from transformers.

Table 4. Ablation studies of each model fine-tuned on different data sets. Italics indicate the best performance.

Models	$Data_{raw}$			$Data_{MR}$[a]			$Data_{DAGA}$[b]		
	P,[c] %	R,[d] %	F1,[e] %	P, %	R, %	F1, %	P, %	R, %	F1, %
Gated recurrent units	94.92	93.04	93.97	95.68	94.2	94.94	94.64	94.59	94.61
BiLSTM[f]	97.12	95.99	96.55	97.72	97.15	97.43	97.14	96.86	97
Base BERT[g]	*98.55*	98.7	98.63	98.25	98.6	98.43	98.6	98.5	98.55
Chinese-BERT-wwm	98.35	98.5	98.43	98.5	98.7	98.6	98.45	98.7	98.58
Chinese-BERT-wwm-ext	98.4	98.5	98.45	98.11	98.7	98.4	98.8	98.9	98.85
Chinese-BERT-base	82.92	85.36	84.12	88.37	88.88	88.63	94.42	95.7	95.06
Chinese-BERT-large	95.42	95.7	95.56	94.95	96.42	95.68	97.53	97.25	97.39
PCL-MedBERT	98.37	99.08	98.72	98.18	98.89	98.53	98.7	*99.23*	98.96
PCL-MedBERT-wwm	98.42	*99.18*	*98.8*	*98.51*	98.99	*98.75*	*98.94*	99.13	*99.03*
Our model	97.84	98.6	98.22	98.32	*99.03*	98.68	98.18	99.08	98.6

[a]MR: mention replacement.

[b]DAGA: data augmentation with a generation approach.

[c]P: precision.

[d]R: recall.

[e]F1: F1 score.

[f]BiLSTM: bidirectional long short-term memory.

[g]BERT: bidirectional encoder representations from transformers.

Table 5. Performance comparison of our model on various entity types after fine-tuning our model with different data sets. Italics indicate the best performance.

Methods	PER[a]			LOC[b]			ORG[c]			DAT[d]		
	P,[e] %	R,[f] %	F1,[g] %	P, %	R, %	F1, %	P, %	R, %	F1, %	P, %	R, %	F1, %
$Data_{raw}$	99.21	99.52	99.36	96.15	95.24	95.69	97.06	98.02	97.54	97.42	98.55	97.98
$Data_{DAGA}$[h]	99.37	*99.84*	99.6	95.28	96.19	95.73	96.43	98.02	97.22	98.27	99.23	98.75
$Data_{MR}$[i]	99.36	99.36	99.36	94.44	97.14	95.77	96.1	97.69	96.89	*98.75*	*99.42*	*99.08*
$Data_{DAGA+MR}$	*99.84*	99.68	*99.76*	*96.23*	*97.14*	*96.68*	*97.39*	*98.68*	*98.03*	98.65	99.13	98.89

[a]PER: personal name.

[b]LOC: location.

[c]ORG: organization name.

[d]DAT: date.

[e]P: precision.

[f]R: recall.

[g]F1: F1 score.

[h]DAGA: data augmentation with a generation approach.

[i]MR: mention replacement.

Table 6. Symbols and meanings of additionally built training sets.

Symbols	Meaning
	Randomly selected sample comprising 10% of $data_{raw}$.
	Randomly selected sample comprising 50% of $data_{raw}$.
a,b	Mixed data from and the entire data set generated by DAGA and MR.
	Mixed data from and randomly selected data generated by DAGA and MR.

[a]DAGA: data augmentation with a generation approach.

[b]MR: mention replacement.

Table 7. Results of TinyBERT after fine-tuning on different data volumes.

Data Volume	P,[a]%	R,[b]%	F1,[c]%
	91.33	95.26	93.26
	97.46	98.36	97.91
d,e	98.13	98.89	98.51
	98.51	99.08	98.8

[a]P: precision.

[b]R: recall.

[c]F1: F1 score.

[d]DAGA: data augmentation with a generation approach.

[e]MR: mention replacement.

Table 8. Efficiency comparison of the benchmark models.

Models	CPU[a] time, seconds	Difference vs our model, %	GPU[b] time, seconds	Difference vs our model, %
Gated recurrent units	100.76	–36.31	56.45	–9.52
BiLSTM[c]	98.61	–37.68	54.94	–11.94
Base BERT[d]	262.81	39.8	78.02	20.03
Chinese-BERT-wwm	259.96	39.16	78.07	20.08
Chinese-BERT-wwm-ext	263.23	39.89	77.64	19.64
Chinese-BERT-base	220.93	28.38	76.28	18.21
Chinese-BERT-large	698.99	77.36	117.05	46.7
PCL-MedBERT	261.53	39.5	76.44	18.38
PCL-MedBERT-wwm	260.38	39.23	78.02	20.03
Our model	158.22	N/A[e]	62.39	N/A

[a]CPU: central processing unit.

[b]GPU: graphics processing unit.

[c]BiLSTM: bidirectional long short-term memory.

[d]BERT: bidirectional encoder representations from transformers.

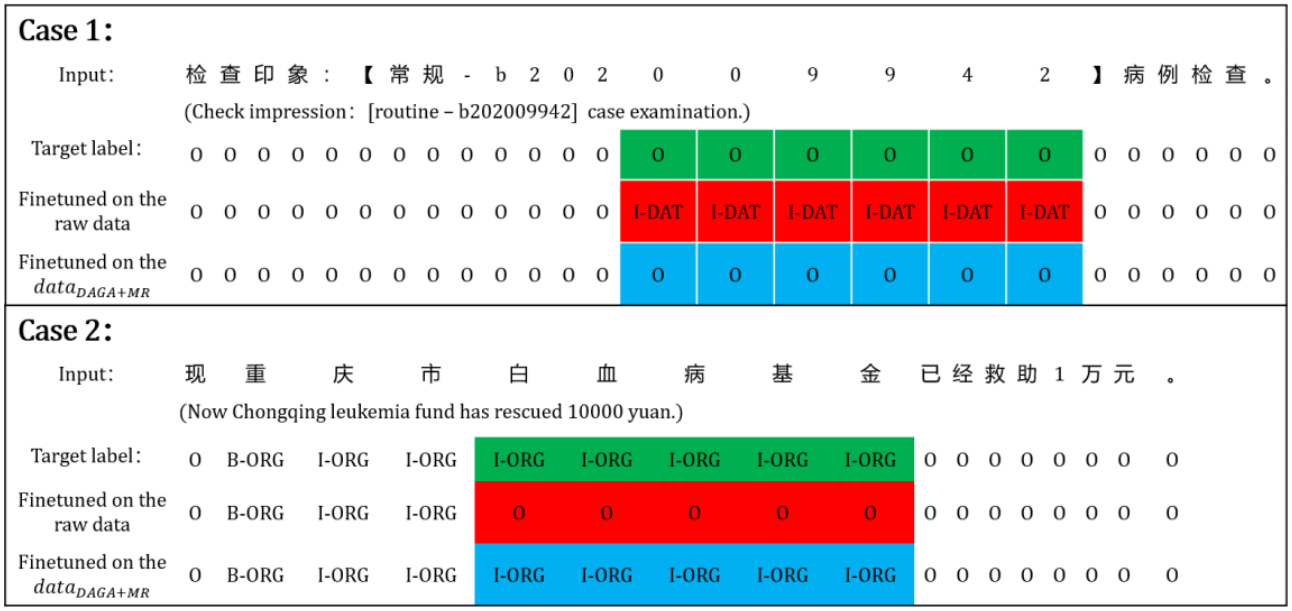
[e]N/A: not applicable.

Case Studies

To visually verify the effectiveness of our proposed method, we used case studies as examples, as shown in Figure 4. In case 1, our model incorrectly classified the number “009942” from the “O” type as the DAT type after fine-tuning on the raw data. This was corrected after fine-tuning on our hybrid augmented data set. In case 2, the entity “白血病基金” (leukemia fund), which should have the ORG type, was not recognized when our

model was fine-tuned on the raw data. However, our model was able to modify this result through context semantics after fine-tuning on the hybrid data set. These case studies demonstrate the effectiveness of the hybrid data augmentation method.

Figure 4. Examples of the results of fine-tuning our model on the hybrid augmented data set. DAT: date; ORG; organization name; DAGA: data augmentation with a generation approach; MR: mention replacement.



Discussion

The main contributions of this paper are to (1) describe a new and efficient model that incorporates a TinyBERT and a CRF model to deidentify PHI in Chinese EHRs; (2) describe a hybrid data augmentation method utilizing a sentence generation strategy and an MR strategy for enhancing Chinese EHRs; and (3) report that our proposed method surpasses other baseline methods on both performance and efficiency. This could be for two possible reasons. First, the attention mechanism of TinyBERT and the optimal searching strategy of the CRF model ensured that our model learned the global features of texts well, and the lightweight parameters kept it from overfitting in the training process. Second, the DAGA generated more training data with more diversity and less noise for increasing the prior knowledge of data distribution for learning. The MR strategy randomly replaced entities in a sentence for learning representations of entities from diverse perspectives, which provided richer contextual information. The worse performance a model had after fine-tuning on raw data, the greater the performance improvement it could obtain after fine-tuning on the hybrid augmented data set. Additionally, the training curves of our model on $data_{raw}$ and $data_{DAGA+MR}$ are shown in Figure 5. This shows that our model quickly converged during training, which greatly reduced the training cost.

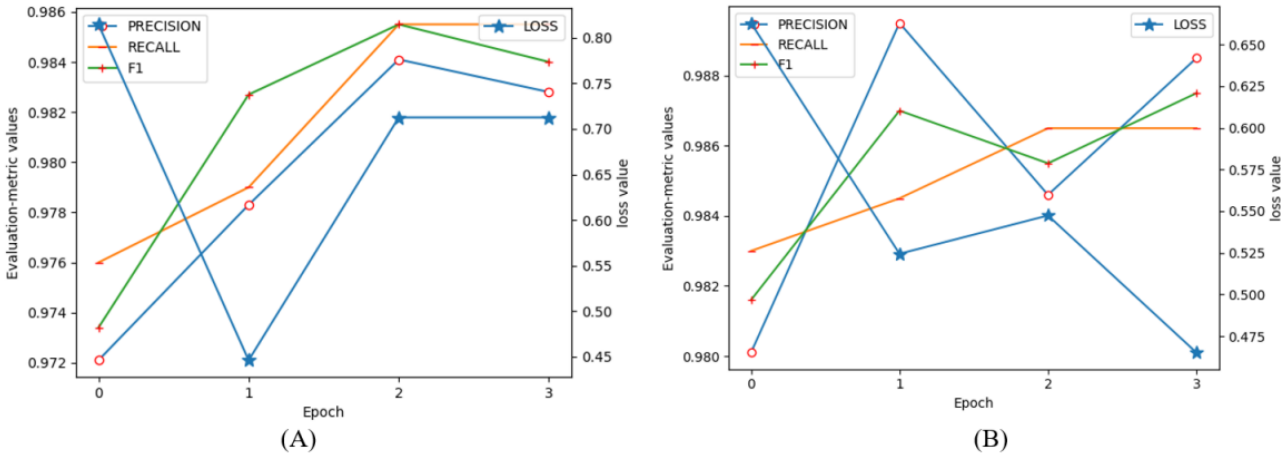
In addition, we performed an analysis to determine why the performance of some baseline methods degraded after fine-tuning on $data_{raw}$ or $data_{DAGA}$. We found that, on the one hand, there may have been disadvantages arising from the data sparsity of $data_{DAGA}$, which hampered the ability of the models to focus on useful contextual semantic information in sentences, impairing feature extraction. On the other hand, applying the MR strategy to the raw data set tended to generate duplicate data, which could have resulted in overfitting in the training process. These 2 shortcomings had a greater impact on the Chinese-BERT-large model, because that model has more transformers and parameters, and is therefore more sensitive to data disturbances [25]. However, the hybrid augmented data could not ensure that our model improved its performance on each type of entity, although the performance on the overall test set was still improved. Moreover, the pretraining data set had a great impact on the downstream tasks. Though the base BERT was pretrained on the English corpus, it obtained much better performance than Chinese-BERT-base. Chinese-BERT-base and Chinese-BERT-large were pretrained with Chinese word information including words, glyphs, and pinyin information. In our experiments, we solely fine-tuned these models with word information, like the other BERT-based models, and found that this led to heavy performance degradation on $data_{raw}$.

The input of our proposed model is structured data, which needs to be correctly prepared from the raw collected data. Although we employed a BERT model to improve the efficiency of the annotating process, the generalizability of this method to EHRs in different languages has not been proved. Furthermore, the location information could have correlations with disease type, although we did not specifically evaluate the influence of PHI deidentification on clinical data mining in this paper.

This paper proposes an efficient and effective model that integrates a TinyBERT and a CRF model for the task of deidentifying PHI in Chinese EHRs. This model relieves the high dependency on computing resources of previous models and improves the efficiency of the task. To overcome the limitation of insufficient annotated data, we propose a hybrid data augmentation method, which uses a generation approach and an MR strategy to create a new data set for fine-tuning the model. Our experimental results demonstrate that the

performance of our model was greater than baseline models and also had the highest efficiency of all the experimental benchmark models.

Figure 5. Training curves of our model on (A) the raw data set and (B) the hybrid augmented data set.



Acknowledgments

The project was supported by National Key R&D Program of China (grants 2018YFC0116702 and 2018YFB2101204).

Data Availability

The data sets used and analyzed during the current study are available from the first author upon reasonable request.

Authors' Contributions

PW, YongL, LY, and CW led the application of the method, conducted the experiments, and analyzed the results. SL, LL, ZZ, SL, FW, HW, and YingL participated in the data extraction and preprocessing. YongL and CW participated in manuscript revision. LL provided theoretical guidance and revised the manuscript.

Conflicts of Interest

None declared.

References

1. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform 2015 Dec;58 Suppl:S20-S29 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.020] [Medline: 26319540]
2. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. J Biomed Inform 2017 Nov;75S:S4-S18 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.011] [Medline: 28614702]
3. Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple M. Identifying personal health information using support vector machines. 2006 Presented at: Proceedings of i2b2 workshop on challenges in natural language processing for clinical data; Nov 10-11, 2006; Washington, DC p. 10-11.
4. Thomas SM, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp 2002:777-781 [FREE Full text] [Medline: 12463930]
5. Neamatullah I, Douglass MM, Lehman LWH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008 Jul 24;8(1):32-17 [FREE Full text] [doi: 10.1186/1472-6947-8-32] [Medline: 18652655]
6. Zhao Z, Yang M, Tang B, Zhao T. Re-examination of Rule-Based Methods in Deidentification of Electronic Health Records: Algorithm Development and Validation. JMIR Med Inform 2020 Apr 30;8(4):e17622 [FREE Full text] [doi: 10.2196/17622] [Medline: 32352384]
7. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004 Feb;121(2):176-186. [doi: 10.1309/E6K3-3GBP-E5C2-7FYU] [Medline: 14983930]

8. Li D, Kipper-Schuler K, Savova G. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. 2008 Presented at: Proceedings of the workshop on current trends in biomedical natural language processing; Jun 19-24, 2008; Columbus, OH. [doi: 10.3115/1572306.1572326]
9. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5):550-563 [FREE Full text] [doi: 10.1197/jamia.M2444] [Medline: 17600094]
10. Li M, Scaiano M, El Emam K, Malin BA. Efficient Active Learning for Electronic Medical Record De-identification. AMIA Jt Summits Transl Sci Proc 2019;2019:462-471 [FREE Full text] [Medline: 31259000]
11. Jian Z, Guo X, Liu S, Ma H, Zhang S, Zhang R, et al. A cascaded approach for Chinese clinical text de-identification with less annotation effort. J Biomed Inform 2017 Sep;73:76-83 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.017] [Medline: 28756160]
12. Du L, Xia C, Deng Z, Lu G, Xia S, Ma J. A machine learning based approach to identify protected health information in Chinese clinical text. Int J Med Inform 2018 Aug;116:24-32. [doi: 10.1016/j.ijmedinf.2018.05.010] [Medline: 29887232]
13. 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]
14. Misawa S, Taniguchi M, Miura Y, Ohkuma T. Character-based bidirectional LSTM-CRF with wordscharacters for Japanese named entity recognition. 2017 Presented at: Proceedings of the 1st Workshop on SubwordCharacter Level Models in NLP; Sep 7, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/w17-4114]
15. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017 May 01;24(3):596-606 [FREE Full text] [doi: 10.1093/jamia/ocw156] [Medline: 28040687]
16. Zhang Y, Gan Z, Fan K, Chen Z, Henao R, Shen D, et al. Adversarial feature matching for text generation. 2017 Presented at: Proceedings of the 34th International Conference on Machine Learning; Aug 6-11, 2017; Sydney, Australia.
17. Jiawei W, Xin W, Wang WY. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, MN. [doi: 10.18653/v1/n19-1120]
18. Festag S, Spreckelsen C. Privacy-Preserving Deep Learning for the Detection of Protected Health Information in Real-World Data: Comparative Evaluation. JMIR Form Res 2020 May 05;4(5):e14064 [FREE Full text] [doi: 10.2196/14064] [Medline: 32369025]
19. Tang B, Jiang D, Chen Q, Wang X, Yan J, Shen Y. De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models. AMIA Annu Symp Proc 2019;2019:857-863 [FREE Full text] [Medline: 32308882]
20. Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. arXiv Preprint posted online May 29, 2016.
21. Zheng L, Guha N, Anderson BR, Henderson P, Ho DE. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. 2021 Presented at: Proceedings of the 18th International Conference on Artificial Intelligence and Law; Jun 21-25, 2021; São Paulo Brazil. [doi: 10.1145/3462757.3466088]
22. Xiaoqi J, Yichun Y, Lifeng S, Xin J, Xiao C, Linlin L, et al. TinyBERT: Distilling BERT for natural language understanding. 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; Nov 16-20, 2020; Punta Cana, Dominican Republic. [doi: 10.18653/v1/2020.findings-emnlp.372]
23. Bosheng D, Linlin L, Lidong B, Canasai K, Thien HN, Shafiq J, et al. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; Nov 16-20, 2020; Punta Cana, Dominican Republic. [doi: 10.18653/v1/2020.emnlp-main.488]
24. Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 2019 Presented at: Proceedings of Empirical Methods in Natural Language Processing; Oct 3-7, 2019; Hong Kong, China.
25. Sutton C, McCallum A. An Introduction to Conditional Random Fields. In: Foundations and Trends in Machine Learning. Norwell, MA: Now Publishers; 2012:267-373.
26. Devlin J, MingWei C, Kenton L, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, MN.
27. Dai X, Adel H. An analysis of simple data augmentation for named entity recognition. 2020 Presented at: Proceedings of the 28th International Conference on Computational Linguistics; Dec 8-13, 2020; Barcelona, Spain. [doi: 10.18653/v1/2020.coling-main.343]
28. Culotta A, McCallum A. Reducing labeling effort for structured prediction tasks. 2005 Presented at: Proceedings of the 25th Conference of Association for the Advancement of Artificial Intelligence; Jul 9-13, 2005; Pittsburgh, PA.
29. Lughofer E. Hybrid active learning for reducing the annotation effort of operators in classification systems. Pattern Recognit 2012 Feb;45(2):884-896. [doi: 10.1016/j.patcog.2011.08.009]
30. Zhiheng H, Wei X, Kai Y. Bidirectional LSTM-CRF models for sequence tagging. arXiv Preprint posted online Aug 9, 2015.

31. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014 Presented at: Proceedings of the 14th Empirical Methods in Natural Language Processing; Oct 25-29, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1179]
32. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-Training With Whole Word Masking for Chinese BERT. IEEE/ACM Trans Audio Speech Lang Process 2021;29:3504-3514. [doi: 10.1109/taslp.2021.3124365]
33. Zijun S, Xiaoya L, Xiaofei S, Yuxian M, Xiang A, Qing H, et al. ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information. 2021 Presented at: the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; Aug 8-14, 2021; Bangkok, Thailand. [doi: 10.18653/v1/2021.acl-long.161]
34. PCL-MedBERT. Pengcheng Laboratory. URL: https://code.ihub.org.cn/projects/1775 [accessed 2022-08-09]

Abbreviations

BERT: bidirectional encoder representations from transformers
CPU: central processing unit
CRF: conditional random field
DAGA: data augmentation with a generation approach
DAT: date
EHR: electronic health record
GPU: graphics processing unit
GRU: gated recurrent units
KD: knowledge distillation
LOC: location
LSTM: long short-term memory
MR: mention replacement
NER: named entity recognition
ORG: organization name
PER: personal name
PHI: protected health information
RNN: recurrent neural network

Edited by C Lovis; submitted 21.03.22; peer-reviewed by X Li, Y Liu, Z Li; comments to author 25.05.22; revised version received 19.07.22; accepted 31.07.22; published 30.08.22.

Please cite as:
Wang P, Li Y, Yang L, Li S, Li L, Zhao Z, Long S, Wang F, Wang H, Li Y, Wang C
An Efficient Method for Deidentifying Protected Health Information in Chinese Electronic Health Records: Algorithm Development and Validation
JMIR Med Inform 2022;10(8):e38154
URL: https://medinform.jmir.org/2022/8/e38154
doi:10.2196/38154
PMID:36040774

Original Paper

An mHealth-Based Health Management Information System Among Health Workers in Volta and Eastern Regions of Ghana: Pre-Post Comparison Analysis

Young-ji Lee[1], MDP; Seohyun Lee[2], PhD; SeYeon Kim[1], MPH; Wonil Choi[1], MPH; Yoojin Jeong[1], MPH; Nina Jin Joo Rhim[3], MPH; Ilwon Seo[4], BS; Sun-Young Kim[1,5], PhD

[1]Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Gwanak Campus, Seoul, Republic of Korea

[2]Department of Global Public Administration, Yonsei University Mirae Campus, Wonju, Republic of Korea

[3]Good Neighbors International, Seoul, Republic of Korea

[4]Good Neighbors Ghana, Accra, Ghana

[5]Institute of Health and Environment, Seoul National University, Gwanak Campus, Seoul, Republic of Korea

Corresponding Author:
Sun-Young Kim, PhD
Department of Public Health Sciences, Graduate School of Public Health
Seoul National University
Gwanak Campus
1 Gwanak-ro
Gwanak-gu
Seoul, 08826
Republic of Korea
Phone: 82 28802768
Email: sykim22@snu.ac.kr

Abstract

Background: Despite the increasing attention to electronic health management information systems (HMISs) in global health, most African countries still depend on inefficient paper-based systems. Good Neighbors International and Evaluate 4 Health have recently supported the Ghana Health Service on the rollout of a mobile health–based HMIS called the *e-Tracker* system in 2 regions in Ghana. The e-Tracker is an Android-based tracker capture app that electronically manages maternal and child health (MCH) data. The Ghana Health Service has implemented this new system in Community Health Planning and Services in the 2 regions (Volta and Eastern).

Objective: This study aims to evaluate changes in health workers' capacity and behavior after using the e-Tracker to deliver MCH services. Specifically, the study assesses the changes in knowledge, attitude, and practice (KAP) of the health workers toward the e-Tracker system by comparing the pre- and postsurvey results.

Methods: The KAP of frontline health workers was measured through self-administered surveys before and after using the e-Tracker system to assess their capacity and behavioral change toward the system. A total of 1124 health workers from the Volta and Eastern regions responded to the pre-post surveys. This study conducted the McNemar chi-square test and Wilcoxon signed-rank test for a pre-post comparison analysis. In addition, random-effects ordered logistic regression analysis and random-effects panel analysis were conducted to identify factors associated with KAP level.

Results: The pre-post comparison analysis showed significant improvement in health workers' capacity, with higher knowledge and practice levels after using the e-Tracker system. As for *knowledge*, there was a 9.9%-point increase (from 559/1109, 50.41% to 669/1109, 60.32%) in the proportion of the respondents who were able to generate basic statistics on the number of children born in a random month within 30 minutes. In the *practice* section, the percentage of respondents who had *scheduled clientencounters* increased from 91.41% (968/1059) to 97.83% (1036/1059). By contrast, responses to the *attitude* (acceptability) became less favorable after experiencing the actual system. For instance, 48.53% (544/1121) initially expressed their preferences for an electronic system; however, the proportion decreased to 33.45% (375/1121) after the intervention. Random-effects ordered logistic regression showed that *days of overwork* were significantly associated with health workers' attitudes toward the e-Tracker system.

Conclusions: This study provides empirical evidence that the e-Tracker system is conducive to enhancing capacity in MCH data management for providing necessary MCH services. However, the change in attitude implies that the users appear to feel less comfortable using the new system. As Ghana plans to scale up the electronic HMIS system using the e-Tracker to the national level, strategies to enhance health workers' attitudes are necessary to sustain this new system.

(JMIR Med Inform 2022;10(8):e29431) doi:10.2196/29431

KEYWORDS

mobile health; mHealth; e-Tracker; health information system; HIS; health information management system; HIMS; District Health Information Management System; DHIMS; maternal and child health; MCH; electronic health record; EHR; health workers

Introduction

Background

A health management information system (HMIS) is a critical component of the health system. According to the World Health Organization, a well-functioning HMIS should "ensure the production, analysis, dissemination, and use of reliable and timely information on health determinants, health systems performance, and health status" [1]. In other words, the key functions of HMIS include the generation, compilation, analysis, synthesis, communication, and use of health information [2]. Among these functions of HMIS, generating health data is particularly crucial as it adds value by providing insights into clinical decision-making and policy implications.

The systematic generation and management of health data in low- and middle-income countries (LMICs) have specific challenges in multiple health focus areas. For example, even the most essential vital statistics such as birth records or maternal and child health (MCH) service provision statistics have not been tracked systematically in many LMICs. In recent years, at least 15,000 newborns died annually, without official records [3]. Similarly, gaps exist between actual service provision and reported data, which makes it difficult for local governments to identify the unmet needs of health services [4]. Health data management is more challenging in resource-constrained settings as the health records are stored in paper-based charts rather than collected electronically. Health workers in such settings generate basic statistics or aggregate the data from paper-based health records and submit the data in person by visiting upper-level facilities such as district or provincial health offices. This manual process is time consuming and often leads to poor data quality [5-8]. In this context, an electronic HMIS has been recognized as an effective and efficient way of addressing this challenge and bridging the quality gap between health care service provision and data management [5,9,10]. Some African countries have recently attempted to implement mobile health (mHealth)–based HMIS as it can be operated using relatively simple software at a lower cost [11,12]. The use of mobile phones or tablet computers for operating HMIS can also address logistic problems, including limited access to fixed broadband internet [13,14], lack of electricity supply [14-17], and financial and human resource deficits in low-resource settings [9,10,18-22].

Ghana is an LMIC that has adopted an mHealth-based HMIS by implementing the MCH data capture app on a tablet computer. Originally, the HMIS in Ghana was initiated as a purely paper-based system in which all stages of data management, from data collection to storage, were performed manually. Once computers became available, the process started transitioning from paper-based to electronic systems at the district level. In 2012, Ghana Health Service (GHS), an implementing agency under the Ministry of Health, implemented the official health service data management software platform, District Health Information Management System, which enabled district health officers to manage health data electronically. However, lower-level health facilities still maintained a paper-based HMIS, which is highly error prone [14,15]. This transition in Ghana was partial as it did not include peripheral community-level health facilities [4]. Community health facilities in Ghana are called Community-based Health Planning and Services compounds and belong to the lowest level of the public health structure in Ghana [23].

In response to the growing demand for an efficient data management system, the GHS implemented the e-Tracker system in 2015, applying it first to family planning and MCH services at the community level for effective and efficient data management of the services [14,24]. The e-Tracker is an Android version of the individual client-based module in the District Health Information System 2. Developed by Oslo University in 2005, this open-source software platform enables the reporting, analysis, and sharing of data for the public health sector. The system is operated on tablet computers to resolve common obstacles such as limited electricity, internet access, lack of financing, and limited human resources by allowing offline data collection and management via portable devices [8,10,14]. GHS is taking the lead to increase health workers' capacity not only for data recordings but also for managing tasks such as *tracking clients who drop out of care*, *scheduling*, *monitoring health services*, and *generating reports* [14,16]. Throughout these transitions, the goal of the HMIS in Ghana has been to support transparent decision-making for nationwide health sector programs [4,21].

For a successful transition from a paper-based health record to an electronic HMIS, the willingness of end users to change their workflow is essential for the sustained use of a new system. In this light, ensuring health workers' acceptability and positive perceptions of the change in practice is one of the key facilitating factors in implementing the e-Tracker, as health workers in community health facilities are frontline workers responsible for managing health data [9,17]. Thus, health workers' acceptability of this new system is considered a prerequisite for the successful implementation of an mHealth-based HMIS [10]. The study by Zargara et al [14] reported that a new system's realignment of work practices is a determinant of MCH service

provision quality. The study also reported that the key challenges in transitioning from paper-based to electronic health records were "an increase in workload occurred by double work" and "low computer literacy" [4,9]. A working paper published by the US Agency for International Development and Measure Evaluation showed mixed results in that the health workers from the 4 districts in Ghana's Central Region did not use the full functionality of the new mHealth-based HMIS, such as data analysis. However, most of them were satisfied with the advanced technology for managing health data [24].

Objective

To further investigate the frontline health workers' capacity, perceptions, and practice toward the e-Tracker, this study conducted a pre-post survey to measure knowledge, attitude, and practice (KAP) among the health workers at Community-based Health Planning and Services compounds in the Volta and Eastern regions of Ghana where the e-Tracker was gradually rolled out to all districts within the region. The empirical findings of this study are expected to provide grounds and political implications for the national scale-up of the e-Tracker system.

Methods

Study Sample

This study used a quasi-experimental pre- and postanalysis design. The KAP on MCH data management using the e-Tracker was investigated through paper-based pre-post surveys. The study adopted a purposive sampling method, recruiting respondents during the e-Tracker system training sessions in the Volta (recently renamed the Oti and Volta regions) and Eastern regions in Ghana. Although there were no specific inclusion or exclusion criteria for survey participants, the respondents were presumed to possess qualifications to fulfill the research purpose as the eligible participants of the training session were frontline health workers who were in charge of providing health services and managing patient data.

For the presurvey, respondents were recruited during the initial training session of the e-Tracker system, where they were introduced to the system. The postsurvey was conducted during the refresher training after 3 to 10 months of e-Tracker use. A total of 2396 health workers participated in the presurvey; however, only 46.9% (1124/2396) of respondents who had participated in the initial training (ie, the presurvey) were able to rejoin the refresher training (ie, the postsurvey) as the GHS arranged to place a portion of the initial participants with newly employed health workers who had not received training opportunities. As a result, approximately half of the respondents from the presurvey were replaced with newly participating health workers, shrinking the study sample size (respondents who participated in both pre- and postsurveys) to 1124. The final set of respondents comprised different types of community health workers (community health nurses [CHNs] or community health officers [CHOs], midwives, enrolled nurses, and field technicians) working in the Volta and Eastern regions (Multimedia Appendix 1).

Data Collection and the Questionnaire

The survey was conducted between October 2018 and November 2019. It was designed as a paper-based, self-administered questionnaire collected by staff from Good Neighbors International, the implementing partner of the e-Tracker training program. Responses were entered manually into a Microsoft Excel spreadsheet by the research team. The questionnaire comprised 43 multiple-choice and yes or no questions covering the content domains of demographics and KAP (Multimedia Appendix 2).

First, the *knowledge* section of the questionnaire asked respondents whether they could retrieve specific information on MCH statistics within 30 minutes. The 10 tasks listed in the questionnaire were designed based on observations during the field visit. The questions asked about the respondents' perceived capacity to generate basic statistics (such as the number of children born, stillbirths, and women who came for antenatal care visits in a specific month in the catchment area). The first half of the questions were intended to ask whether aggregate data could be generated for a randomly selected month. The remaining 5 questions asked whether health workers could retrieve aggregate data for the month when the survey was conducted. Second, the section for *attitude* comprised 8 questions with a 5-point Likert scale to identify the level of acceptability of using an electronic device for managing MCH records. The questions asked about the respondents' willingness, perception, and preference for using an electronic device for MCH data management. Third, the *practice* section comprised questions on the practice of 8 specific tasks related to MCH data management and the perceived difficulty in performing those tasks. In addition, the use of a tablet computer for MCH data management and the frequency of electronic devices used for MCH data management were asked. As the data on tablet computer use were systematically inaccessible, self-reported responses were used to assess the practice.

Statistical Analysis

Data from the pre- and postsurveys coded in the spreadsheets were imported into STATA (version 14; StataCorp LLC). Unique identifications were randomly generated for each participant, which allowed each participant's pre- and postsurvey variables to be reliably matched. McNemar chi-square and Wilcoxon signed-rank tests were used for pre-post comparison analyses. In addition, to investigate the factors associated with each KAP component, random-effects ordered logistic regression and random-effects panel analysis were conducted. For the dependent variable, a Cronbach α test was performed for each KAP to test internal consistency for aggregating different responses to a single score. The duration of the intervention (ie, use of the tablet-based e-Tracker system in managing MCH data) was selected as the explanatory variable, and the control variables were categorized into enabling environmental, demographic, and working condition factors. The explanatory variable, represented by the "number of days of using the e-Tracker system for MCH data management," varied as the time points for the presurvey (the initial training workshop) and the postsurvey (refresher training) were different across the districts covered. The variable *days of overwork* was

included only in the regression model, as introducing an mHealth-based HMIS may have intensified the health workers' workload, increasing resistance toward the emergent system (Table 1).

Table 1. Analysis framework of regression analysis.

Variables	Description
Dependent variables	
Knowledge	Knowledge of MCH[a] data management (score between 0 and 10)
Attitude	Attitude on using an electronic device to manage MCH data (scaled between 1=most negative and 5=most positive)
Practice	Frequency of using an electronic device to manage MCH data (scaled between 1=never and 5=every time)
Explanatory variable	
Duration of e-Tracker use	Days of using the e-Tracker system via a tablet computer
Control variables	
Environmental factor	Level of internet connection at the health facility
Demographic factors	Age, sex, educational level, working experience, job position, and use of mobile phone
Working condition	Days of overwork

[a]MCH: maternal and child health.

Ethics Approval

This study received ethics approval from the GHS Ethics Review Committee (GHS-ERC009/09/18; Multimedia Appendix 3).

Results

Summary of the Respondents' Characteristics

Table 2 presents descriptive statistics for respondents who participated in the presurvey only (group A) and those who participated in both pre- and postsurveys (group B). The 2 groups of respondents were analyzed to identify any significant differences that might have been caused because of a change in sample size. In group A, approximately 53.23% (676/1272) were from the Volta region, whereas in group B, it was 31.76% (357/1124). Approximately 83.81% (1066/1272) of respondents were female in group A, whereas it was 79.27% (891/1124) for group B. As for education, both groups showed a similar proportion for each academic level; however, group A tended to have a slightly higher educational background. Specifically, the percentages of respondents with diplomas and bachelor's degrees were about 3% points and 2% points higher for group A, respectively. Similarly, group A respondents tended to engage in a higher job position as 28.38% (361/1272) were midwives, whereas 15.84% (178/1124) were midwives in group B. Both groups had a high rate of mobile phone use as >96% indicated the *use of their own mobile phones*. As for internet access, approximately 11.4% (145/1272) and 11.48% (129/1124) of respondents answered that their facilities had no internet access, whereas 31.84% (405/1272) and 31.23% (351/1124) responded with an acceptable level of internet access at work sites groups A and B, respectively. In addition, only 6.21% (79/1272) and 6.14% (69/1124) in groups A and B, respectively, answered that their facilities had very reliable internet access. Regarding the average age and working experience, the average age of group A respondents was approximately 1 year higher than those in group B. The average duration of using an e-Tracker system was 187.55 (51.17) days. The differences between the 2 groups in the chi-square analysis results were statistically significant for all demographic factors. Notably, the results indicated that health workers with a relatively lower educational background and shorter work experience participated in both pre- and postsurveys by rejoining the refresher training.

Table 2. Sociodemographic characteristics of the respondents.

Characteristics	Group A, presurvey only (n=1272)	Group B, matched (n=1124)	*P* value
Region, n (%)			
Volta	676 (53.23)	357 (31.76)	<.001[a]
Eastern	596 (46.86)	767 (68.24)	<.001[a]
Sex, n (%)			
Male	204 (16.04)	233 (20.73)	.004[a]
Female	1066 (83.81)	891 (79.27)	.004[a]
Missing	2 (0.16)	0 (0)	.004[a]
Educational level, n (%)			
Certificate	1051 (82.79)	993 (88.35)	<.001[a]
Diploma	179 (14.13)	121 (10.77)	<.001[a]
Bachelor's degree	37 (2.92)	9 (0.8)	<.001[a]
Master's degree	1 (0.08)	1 (0.09)	<.001[a]
Other	1 (0.08)	0 (0)	<.001[a]
Missing	4 (0.32)	0 (0)	<.001[a]
Job description, n (%)			
Community health nurse or community health officer	901 (70.83)	941 (83.72)	<.001[a]
Enrolled nurse	2 (0.16)	1 (0.09)	<.001[a]
Midwife	361 (28.38)	178 (15.84)	<.001[a]
Field technician	1 (0.08)	2 (0.18)	<.001[a]
Other	3 (0.24)	1 (0.09)	<.001[a]
Missing	4 (0.31)	1 (0.09)	<.001[a]
Use of mobile phone, n (%)			
Yes (use my own mobile phone)	1224 (96.23)	1091 (97.06)	.05[a]
Yes (share a mobile phone with family)	5 (0.39)	0 (0)	.05[a]
No (do not use or have a mobile phone)	2 (0.16)	1 (0.09)	.05[a]
Missing	41 (3.22)	32 (2.85)	.05[a]
Access to the internet, n (%)			
No internet	145 (11.4)	129 (11.48)	.17[a]
Very poor	99 (7.78)	102 (9.07)	.17[a]
Poor	139 (10.93)	113 (10.05)	.17[a]
Acceptable	405 (31.84)	351 (31.23)	.17[a]
Reliable	347 (27.28)	329 (29.27)	.17[a]
Very reliable	79 (6.21)	69 (6.14)	.17[a]
Missing	58 (4.56)	31 (2.76)	.17[a]
Age (years), mean (SD)	32.89 (6.65)	31.45 (5.44)	<.001[b]
Duration of work as a health professional (years), mean (SD)	6.80 (5.74)	5.34 (5.02)	<.001[b]

Characteristics	Group A, presurvey only (n=1272)	Group B, matched (n=1124)	*P* value
Days of using an e-Tracker system, mean (SD)	N/A[c]	187.55 (51.17)	N/A

[a]*P* value derived from chi-square test.

[b]*P* value derived from 1-way ANOVA test.

[c]N/A: not applicable.

Knowledge

The responses were analyzed using the McNemar chi-square test to evaluate the pre- and postlevel knowledge. As shown in Table 3, there were statistically significant improvements for all question items. For example, there was a 9.9%-point increase (from 559/1109, 50.41% to 669/1109, 60.32%) in the proportion of respondents who were able to *generate basic statistics within 30 minutes on the number of children born* for a randomly selected month. In addition, the proportion of respondents who were able to retrieve the *number of pregnant women expected to deliver and those scheduled for their second postnatal care visit* during the month of the survey increased by 8.9% points (from 369/1108, 33.3% to 468/1108, 42.24%) and 8.0% points (from 283/1109, 25.52% to 337/1109, 33.54%), respectively.

After obtaining an aggregated score for the levels of knowledge by summing up the total number of tasks that an individual respondent was capable of, a Cronbach α test was conducted to verify the reliability of the aggregated scores. The scale reliability coefficients of the pre- and postsurvey responses were α=.71 and α=.72, respectively. As the test results showed acceptable reliability, 10 self-reported responses were aggregated into a single score ranging between 0 and 10. A random-effects ordered logistic analysis showed no significant impact of intervention duration on health workers' knowledge (odds ratio [OR] 1.00, 95% CI 0.99-1.00; Table 4). However, respondents' sex, working years, and job positions had a statistically significant association with their level of knowledge. Participants who were female tended to have lower knowledge levels than participants who were male (OR 0.53, 95% CI 0.41-0.70). Moreover, health workers with longer working years had higher knowledge levels (OR 1.06, 95% CI 1.03-1.10), and compared with CHN or CHO, midwives appeared to have higher knowledge levels (OR 2.86, 95% CI 2.03-4.02).

Table 3. Result of pre-post analysis for knowledge (N=1109).

Knowledge on data management	Presurvey	Postsurvey	*P* value[a]
Can retrieve basic statistics on the total number of following items for a random month within 30 minutes, n (%)			
Children born	559 (50.41)	669 (60.32)	<.001
Family planning counseling provided	723 (65.19)	783 (70.6)	.001
Stillbirths	282 (25.43)	354 (31.92)	<.001
Women visiting the facility for postpartum complications	222 (20.02)	272 (24.53)	.003
Women visiting for their first antenatal care	480 (43.28)	544 (49.05)	.001
Can retrieve basic statistics on the total number of following items during the month of the survey within 30 minutes, n (%)			
Defaulters for measles immunization[b]	601 (54.24)	663 (59.84)	.001
Pregnant women who are expected to deliver[b]	369 (33.30)	468 (42.24)	<.001
Children aged <l year	626 (56.45)	665 (59.16)	.05
Women scheduled for their second postnatal care visit	283 (25.52)	377 (33.54)	<.001
Women who are in their first trimester of pregnancy[b]	442 (39.89)	496 (44.77)	.002

[a]*P* value derived from the McNemar chi-square test.

[b]A total of 1108 responses was matched.

Table 4. Result of regression analysis for knowledge.

Characteristics	Odds ratio (95% CI)	SE	*P* value
Days of using the e-Tracker system via tablet computer	1.00 (0.99-1.00)	0.00	.48
Age (years)	0.99 (0.97-1.02)	0.01	.67
Sex (reference: male)	0.53 (0.41-0.70)	0.07	<.001
Education level (reference: certificate)			
Diploma	1.20 (0.87-1.65)	0.20	.27
Bachelor's degree	0.77 (0.26-2.27)	0.42	.63
Master's degree	0.02 (0.00-0.73)	0.04	.03
Other	4.11 (0.12-141.19)	7.42	.43
Working years	1.06 (1.03-1.10)	0.02	.001
Job position (reference: CHN[a] or CHO[b])			
Enrolled nurse	3.54 (0.71-17.64)	2.90	.12
Midwife	2.86 (2.03-4.02)	0.50	<.001
Field technician	0.18 (0.00-16.33)	0.42	.46
Other	1.65 (0.13-21.62)	2.16	.70
Use of mobile phone (reference: use own mobile phone)			
Share mobile phone	0.41 (0.04-4.08)	0.48	.44
Do not use mobile phones	4.09 (0.27-61.14)	5.65	.31
Access to the internet (reference: no internet)			
Very poor	1.30 (0.87-1.96)	0.27	.20
Poor	1.40 (0.93-2.10)	0.29	.10
Acceptable	1.18 (0.84-1.65)	0.20	.34
Reliable	1.15 (0.81-1.63)	0.21	.43
Very reliable	0.78 (0.48-1.28)	0.19	.33

[a]CHN: community health nurse.

[b]CHO: community health officer.

Attitude

The Wilcoxon signed-rank test was conducted to assess the prelevel and postlevel of attitude. The initial results showed that approximately 33.99% (379/1115) were *most willing* to manage electronic MCH records (Table 5). However, the proportion decreased to 21.26% (237/1115), whereas the neutral response increased from 18.03% (201/1115) to 28.43% (317/1115). Regarding the preference for paper-based versus electronic-based management, 48.53% (544/1121) initially expressed their preferences for electronic systems; however, the proportion decreased to 33.45% (375/1121) after the intervention. In contrast, the percentage of respondents indifferent to the 2 options increased from 15.7% (176/1121) to 26.32% (295/1121). Compared with the results of the survey, general ideas on using an electronic system or device became less favorable.

The Cronbach α test was conducted to verify the reliability of the 5-point Likert scale for attitude levels. The scale reliability coefficients of the pre- and postsurvey responses were α=.80 and α=.85, respectively. Given the acceptable Cronbach α test results, each of the 8 answers scoring between 1 and 5 was aggregated and converted into one average value and then analyzed using a random-effect panel analysis. As shown in Table 6, the *duration of using the e-Tracker system* was positively associated with attitude toward electronic MCH data management but to a minor degree (coefficient 0.001; *P* value<.001). On the contrary, *days of overwork* showed a negative relationship with the attitude toward the new system. Regarding demographic factors, female health workers tended to favor the new system less. In addition, health workers with diplomas and bachelor's degrees showed more positive attitudes than those with certificates. In contrast, workers with master's degrees had less favorable attitudes. In terms of job positions, enrolled nurses had less favorable attitudes than CHNs and CHOs. Moreover, health workers who shared mobile phones with their families had less favorable attitudes than those with their own mobile phones, implying that the ownership of personal mobile phones may have equipped the respondents with adaptability to the tablet computer system. Access to the internet was also significantly associated with attitudes toward the new system. Health workers who worked at facilities with *very reliable* internet access had more favorable attitudes than those who did not. In summary, some demographic factors,

such as the ownership of personal mobile phones and access to the internet, demonstrated a larger magnitude of effect on attitude than the duration of e-Tracker use.

Table 5. Result of pre-post analysis for attitude.

Attitude toward electronic data management	Presurvey, n (%)	Postsurvey, n (%)	*P* value[a]
Willing to manage MCH[b] records using an electronic system (n=1115)			
1 (least likely)	30 (2.69)	33 (2.96)	<.001
2	41 (3.68)	78 (7)	<.001
3 (neutral)	201 (18.03)	317 (28.43)	<.001
4	464 (41.61)	450 (40.36)	<.001
5 (most likely)	379 (33.99)	237 (21.26)	<.001
Comfortable with managing electronic MCH records (n=1117)			
1 (very uncomfortable)	28 (2.51)	32 (2.86)	<.001
2	46 (4.12)	106 (9.49)	<.001
3 (neutral)	275 (24.62)	383 (34.29)	<.001
4	497 (44.49)	435 (38.94)	<.001
5 (very comfortable)	271 (24.26)	161 (14.41)	<.001
Using an electronic device for managing MCH records is a good idea (n=1120)			
1 (strongly disagree)	6 (0.54)	16 (1.43)	<.001
2	6 (0.54)	34 (3.04)	<.001
3 (neutral)	145 (12.95)	254 (22.68)	<.001
4	398 (35.54)	424 (37.86)	<.001
5 (strongly agree)	565 (50.45)	392 (35)	<.001
Using an electronic device to enter MCH records is difficult for me (n=1116)			
1 (strongly disagree)	419 (37.54)	371 (33.24)	.70
2	171 (15.23)	212 (19)	.70
3 (neutral)	292 (26.16)	354 (31.72)	.70
4	187 (16.76)	145 (12.99)	.70
5 (strongly agree)	47 (4.21)	34 (3.05)	.70
I prefer using an electronic device to manage MCH records than writing them on paper (n=1121)			
1 (strongly disagree)	25 (2.23)	41 (3.66)	<.001
2	25 (2.23)	59 (5.26)	<.001
3 (neutral)	176 (15.7)	295 (26.32)	<.001
4	351 (31.31)	351 (31.31)	<.001
5 (strongly agree)	544 (48.53)	375 (33.45)	<.001
Using an electronic device to enter MCH records is more convenient than writing on paper (n=1120)			
1 (strongly disagree)	13 (1.16)	35 (3.13)	<.001
2	25 (2.23)	57 (5.09)	<.001
3 (neutral)	183 (16.34)	285 (25.45)	<.001
4	371 (33.13)	369 (32.95)	<.001
5 (strongly agree)	528 (47.14)	374 (33.39)	<.001
Using an electronic device to enter MCH records is more accurate than writing on paper (n=1120)			
1 (strongly disagree)	19 (1.70)	39 (3.48)	<.001
2	22 (1.96)	61 (5.45)	<.001
3 (neutral)	185 (16.52)	307 (27.41)	<.001
4	404 (36.07)	355 (31.7)	<.001

Attitude toward electronic data management	Presurvey, n (%)	Postsurvey, n (%)	*P* value[a]
5 (strongly agree)	490 (43.75)	358 (31.96)	<.001
Using an electronic device to enter MCH records is more effective than writing on paper (n=1117)			
1 (strongly disagree)	17 (1.52)	33 (2.95)	<.001
2	14 (1.25)	60 (5.37)	<.001
3 (neutral)	169 (15.13)	295 (26.41)	<.001
4	415 (37.15)	384 (34.38)	<.001
5 (strongly agree)	502 (44.94)	345 (30.89)	<.001

[a]*P* value derived from Wilcoxon signed-rank test.

[b]MCH: maternal and child health.

Table 6. Result of regression analysis for attitude.

Characteristics	Coefficient	SE	*P* value
Days of using the e-Tracker system via a tablet computer	0.001	0.00	<.001
Days of overwork	−0.01	0.00	.002
Age (years)	0.00	0.00	.58
Sex (reference: male)	−0.29	0.04	<.001
Education level (reference: certificate)			
Diploma	0.10	0.05	.04
Bachelor's degree	0.21	0.08	.01
Master's degree	−0.19	0.05	<.001
Other	−0.34	0.07	<.001
Working years	0.00	0.01	.49
Job position (reference: CHN[a] or CHO[b])			
Enrolled nurse	−0.30	0.13	.02
Midwife	0.04	0.06	.45
Field technician	−0.28	0.22	.21
Other	−0.06	0.38	.88
Use of mobile phone (reference: use own mobile phone)			
Share mobile phone	−0.61	0.09	<.001
Do not use mobile phones	0.25	0.22	.25
Access to the internet (reference: no internet)			
Very poor	−0.09	0.07	.20
Poor	−0.07	0.06	.28
Acceptable	0.02	0.05	.70
Reliable	0.12	0.06	.03
Very reliable	0.35	0.07	<.001

[a]CHN: community health nurse.

[b]CHO: community health officer.

Practice

The McNemar chi-square test was conducted for self-reported use of tablet computers for MCH data management and for 8 specific tasks related to MCH data management, such as recording client demographic data or scheduling appointments (Table 7). In addition, the Wilcoxon signed-rank test was performed to assess changes in perceived difficulty in conducting each task following the adoption of the e-Tracker. As expected, the analysis showed that the use of tablet computers for MCH data management increased from 5% (56/1121) to 81.71% (916/1121). As for the frequency of

electronic device use for MCH data management, most respondents (817/1119, 73.01%) answered that they had *never* used an electronic device during the presurvey; however, 26.99% (302/1119) responded that they use it *every time*, 36.73% (411/1119) for *most of the time*, 29.49% (330/1119) for *sometimes*, and 3.75% (42/1119) for *never* after the intervention (ie, during the postsurvey).

In the case of actual practice on 8 specific tasks related to MCH data management, the percentage of respondents who performed 8 tasks showed statistically significant changes after the adoption of the e-Tracker. For example, the percentage of respondents who had *scheduled client encounters* increased from 91.41% (968/1059) to 97.83% (1036/1059). In addition, the percentage of respondents who had *collected individual data into aggregates for the District Health Information Management System 2* increased from 66.04% (702/1063) to 89.93% (956/1063). When asked if they *have ever used statistical data for making a request to the District Health Office*, the percentage of respondents who answered *yes* increased from 52.28% (591/1106) to 70.02% (787/1106). However, no statistically significant changes were found for the percentages of respondents who *produce reports on MCH*, *following up health care defaulters*, and *generate basic statistics other than monthly reports on MCH*. On the one hand, the percentages of respondents who *produce reports on MCH* or *following up health care defaulters* were >97% for both pre- and postsurveys, indicating that the tasks have generally been manageable for the health care workers regardless of the e-Tracker adoption. In contrast, the percentages of respondents who had *generated basic statistics other than monthly reports on MCH* remained at approximately 78.62% (846/1076) and 77.97% (839/1076) throughout the pre- and postsurveys, respectively. This result may imply the limited use of the data aggregation functionality of the e-Tracker.

In terms of perceived difficulty for the 8 tasks, a statistically significant improvement was observed for all 8 tasks after the implementation of the e-Tracker system. For instance, 27.94% (292/1069) responded that *following up with health care defaulters* was *very difficult* before the intervention. However, after using the e-Tracker system, only 6.89% (72/1069) answered that the task was *very difficult*. Moreover, those who found the task *very easy* increased from 7.56% (79/1069) to 15.31% (160/1069).

Unlike the *knowledge* and *attitude* sections, responses from the *practice* section failed to fulfill the acceptable standard through the Cronbach α test. Thus, the practice level for regression analysis was defined as a 5-point Likert scale of the frequency of electronic device use for MCH data management, which was analyzed with random-effects ordered logistic analysis (Table 8). The results showed that health workers with diplomas (OR 1.31, 95% CI 1.02-1.67) had higher practice levels than workers with a certificate educational level. Moreover, respondents with more work experience (OR 1.06, 95% CI 1.03-1.09) tended to show higher practice levels. In the case of environmental factors, internet accessibility was associated with practice level; that is, poor (OR 1.37, 95% CI 0.97-1.93), acceptable (OR 1.61, 95% CI 1.22-2.14), and reliable (OR 1.31, 95% CI 0.98-1.76) internet access showed higher odds than no internet access.

Table 7. Result of pre-post analysis for practice.

Practice on MCH[a] data management	Presurvey, n (%)	Postsurvey, n (%)	*P* value
Use a tablet computer for MCH data management (n=1121)	56 (5)	916 (81.71)	<.001[b]
Frequency of electronic device use for MCH data management (n=1119)			
Every time	15 (1.34)	302 (26.99)	<.001[c]
Most of the time	49 (4.38)	411 (36.73)	<.001[c]
Sometimes	163 (14.57)	330 (29.49)	<.001[c]
Rarely	75 (6.7)	34 (3.04)	<.001[c]
Never	817 (73.01)	42 (3.75)	<.001[c]
The number of respondents who perform the following tasks and the perceived task difficulty			
Recording client demographic data (n=1080)	1028 (95.19)	1062 (98.33)	<.001[b]
Perceived task difficulty			
1 (very difficult)	137 (13.33)	44 (4.14)	<.001[c]
2	179 (17.41)	91 (8.57)	<.001[c]
3	401 (39.01)	395 (37.19)	<.001[c]
4	222 (21.6)	321 (30.23)	<.001[c]
5 (very easy)	103 (10.02)	191 (71.98)	<.001[c]
Scheduling client encounters (n=1059)	968 (91.41)	1036 (97.83)	<.001[b]
Perceived task difficulty			
1 (very difficult)	72 (7.44)	25 (2.41)	<.001[c]
2	153 (15.81)	77 (7.43)	<.001[c]
3	365 (37.71)	349 (33.69)	<.001[c]
4	275 (28.41)	351 (33.88)	<.001[c]
5 (very easy)	122 (12.6)	185 (17.86)	<.001[c]
Tracking client progress over time (n=1064)	992 (93.23)	1027 (96.52)	<.001[b]
Perceived task difficulty			
1 (very difficult)	204 (20.56)	68 (6.62)	<.001[c]
2	211 (21.27)	124 (12.07)	<.001[c]
3	315 (31.75)	372 (36.22)	<.001[c]
4	215 (21.67)	319 (31.06)	<.001[c]
5 (very easy)	54 (5.44)	116 (11.3)	<.001[c]
Following up health care defaulters (n=1069)	1045 (97.75)	1045 (97.75)	.88[b]
Perceived task difficulty			
1 (very difficult)	292 (27.94)	72 (6.89)	<.001[c]
2	264 (25.26)	132 (12.63)	<.001[c]
3	275 (26.32)	349 (33.4)	<.001[c]
4	154 (14.74)	351 (33.59)	<.001[c]
5 (very easy)	79 (7.56)	160 (15.31)	<.001[c]

Practice on MCH[a] data management	Presurvey, n (%)	Postsurvey, n (%)	*P* value
Collecting individual data into aggregates for the District Health Information Management System 2 (n=1063)	702 (66.04)	956 (89.93)	<.001[b]
Perceived task difficulty			
1 (very difficult)	152 (15.70)	45 (4.34)	<.001[c]
2	161 (16.63)	104 (10.04)	<.001[c]
3	213 (22)	277 (26.74)	<.001[c]
4	120 (12.4)	195 (18.82)	<.001[c]
5 (very easy)	55 (5.68)	80 (7.72)	<.001[c]
Producing reports on MCH (n=1088)	1061 (97.52)	1066 (97.98)	.30[b]
Perceived task difficulty			
1 (very difficult)	129 (13.33)	68 (6.56)	<.001[c]
2	215 (22.21)	98 (9.46)	<.001[c]
3	375 (38.74)	403 (38.9)	<.001[c]
4	241 (24.9)	397 (38.32)	<.001[c]
5 (very easy)	103 (10.64)	135 (13.03)	<.001[c]
Generating basic statistics other than monthly reports on MCH (n=1076)	846 (78.62)	839 (77.97)	.70[b]
Perceived task difficulty			
1 (very difficult)	92 (9.50)	41 (3.96)	<.001[c]
2	153 (15.81)	63 (6.08)	<.001[c]
3	264 (27.27)	315 (30.41)	<.001[c]
4	124 (12.81)	190 (18.34)	<.001[c]
5 (very easy)	33 (3.41)	57 (5.5)	<.001[c]
Ever used statistical data for making a request to the District Health Office (n=1106)	591 (52.28)	787 (70.02)	<.001[b]

[a]MCH: maternal and child health.

[b]*P* value derived from the McNemar chi-square test.

[c]*P* value derived from the Wilcoxon signed-rank test.

Table 8. Result of regression analysis for practice.

Practice	Odds ratio (95% CI)	SE	*P* value
Days of using the e-Tracker system via tablet computer	1.00 (1.001-1.004)	0.00	.002
Age (years)	0.98 (0.95-1.00)	0.01	.04
Sex (reference: male)	0.83 (0.68-1.01)	0.08	.06
Education level[a] (reference: certificate)			
Diploma	1.31 (1.02-1.67)	0.16	.03
Bachelor's degree	0.64 (0.27-1.51)	0.28	.31
Master's degree	0.98 (0.13-7.56)	1.02	.98
Other[a]	0	0	0
Working years	1.06 (1.03-1.09)	0.01	<.001
Working position (reference: CHN[b] or CHO[c])			
Enrolled nurse	2.59 (0.70-9.53)	1.72	.15
Midwife	0.92 (0.71-1.18)	0.12	.50
Field technician	0.00 (0.00-0.00)	0.00	.99
Other	1.29 (0.16-10.18)	1.36	.81
Use of mobile phone (reference: use own mobile phone)			
Share mobile phone	2.98 (0.44-20.19)	2.91	.26
Do not use mobile phones	1.22 (0.14-10.43)	1.33	.86
Access to the internet (reference: no internet)			
Very poor	1.21 (0.86-1.70)	0.21	.29
Poor	1.37 (0.97-1.93)	0.24	.07
Acceptable	1.61 (1.22-2.14)	0.23	.001
Reliable	1.31 (0.98-1.76)	0.19	.07
Very reliable	1.11 (0.605-0.74)	0.23	.52

[a]The subcategory of *Other* was removed because of a low number of observations.

[b]CHN: community health nurse.

[c]CHO: community health officer.

Discussion

Principal Findings

This study is the first empirical analysis to explore the change in the KAP of health workers in managing MCH data using the e-Tracker system in Ghana. The pre-post comparison analysis results showed a statistically significant improvement in health workers' knowledge and practice levels of MCH data management. Regarding *knowledge*, the proportion of respondents who reported that they could *retrieve basic MCH statistics* increased after using the e-Tracker system. The changes in the practice level were notable in that there were statistically significant increases in the number of health workers engaging in 8 MCH data management tasks, such as scheduling patients' encounters and tracking patients' progress. Furthermore, a significant improvement was observed in the perceived difficulty of performing these 8 tasks. These results were confirmed by a previous study that reported amelioration in the quality of newborn care of health workers in Malawi after using an mHealth solution called NeoTree [3]. In the case of *attitude*, the level remained positive after using the e-Tracker, which was in line with a previous study that identified high satisfaction with e-Tracker use [24]. However, compared with the results from the presurvey, general ideas on using an electronic system or device became less favorable after experiencing the actual system. An additional regression analysis found that the duration of the intervention (days of using a tablet computer) was positively associated with attitude and practice but to a minor degree. Most importantly, the *days of overwork* showed a statistically significant correlation with attitude level, implying the negative impact of increased workload on health workers' acceptability. This can be explained by the concurrent use of the traditional manual and the new e-Tracker system, which created extra work for health workers, affecting their attitude toward the system. A previous study also identified the realignment of work practice and increased workload because of the introduction of the new system [4]. Furthermore, an environmental factor such as access to the internet was also an essential condition as health workers who worked at facilities with relatively more reliable internet access had more favorable attitudes and higher practice levels. This was confirmed by

previous studies, which ascertained limited access to fixed broadband internet [13,14] and lack of electricity supply [14-17] as obstacles to implementing an electronic HMIS.

Limitations

Despite its contributions in providing empirical evidence on KAP for the new technology, this study has several limitations. First, the results of this study are not free of external validity issues. The participants of this study were limited to health workers at community health facilities in Ghana, and the surveys were conducted during the training sessions for the e-Tracker adoption. Furthermore, the unexpected change in participants during the refresher training reduced the sample size, as only 46.9% (1124/2396) of the presurvey respondents responded to the postsurvey. The major problem was the demographic description of the 2 groups, which showed a statistically significant difference for every demographic factor. This implied that those who participated in both presurveys and postsurveys tended to be less experienced, which could have affected the results. Second, this study failed to establish a complete study environment to compare before and after the e-Tracker system because of the concurrent use of traditional paper-based and new electronic methods during the study period. Such a dual system caused a double burden for data management tasks on health workers, which was presumed to be the cause of less favorable responses in the *attitude* section. This was supported by the regression analysis, which found that the *days of overwork* had a negative association with the overall attitude toward the electronic-based system. Finally, this study focused on quantitative analysis and did not identify the contextual factors that could be captured through in-depth interviews. Thus, further assessment is necessary to understand the complex reasons behind the reluctance or preference for the new system.

Policy Implications

Nevertheless, our study provides insights for drawing policy recommendations to settle the mHealth-based HMIS in Ghana. The findings warrant the benefits of the e-Tracker system, an enhancement in health workers' capacity for MCH data management, which provides justification for the scale-up of the system. To achieve a successful adaptation of the new system, it is necessary to establish national, regional, and facility-level strategies to address users' acceptability. First, ensuring health workers' acceptability is pivotal for the sustained use of the advanced system [9,17]. Previous studies have concluded that double work is one of the challenges of the e-Tracker [4,9,21]. Thus, GHS needs to spur the complete replacement of manual-based data management with the e-Tracker system to enhance job efficiency by reducing the double burden at the national level. Moreover, an effort to develop the infrastructure and environment of community health facilities to secure stable internet access is necessary. Second, on-site training for health workers to use the system should be arranged regularly by the District Health Offices. A previous qualitative study on health workers' perceptions reported that workers who were more accustomed to mobile technology tended to have a positive attitude toward an mHealth system [18]. Other studies have also reported *low computer literacy* as one of the key challenges in transitioning from paper-based to electronic health records [4,9]. Thus, training health workers in data management, defined as collecting, recording, analyzing, and reporting health data, is crucial for more accurate and reliable information and sustained system use [19,25,26]. Finally, facilitative supervision and organizational management are essential to increase users' perceived ease and realign health workers' tasks, which are detrimental to the sustained use of the e-Tracker system [24].

Conclusions

Strengthening the HMIS is vital for improving health outcomes, as it facilitates communication within the health system and contributes to sound and evidence-based decision-making in health policy. However, many low-income countries rely on manual-based HMIS, which has many limitations for collecting and managing health data. The introduction of the e-Tracker, an mHealth-based HMIS, is expected to be an innovative attempt to bridge the gap between existing technology and the outdated practice of paper-based health data management. Currently, there are ongoing efforts to scale up the e-Tracker system nationally in Ghana. This context warrants an increased need to evaluate the new system's effectiveness and sustainability by exploring health workers' capacity and behavioral changes in using the e-Tracker system. The findings of this study will contribute to the successful adoption of the e-Tracker system at the national level by providing grounds for national scale-up and schemes to enhance the sustainability of the system.

Acknowledgments

This study was funded by Samsung Electronics and was supported by the Community Chest of Korea.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Number of respondents by region and district.
[DOCX File , 16 KB - medinform_v10i8e29431_app1.docx]

Multimedia Appendix 2
Evaluation of the mobile health program questionnaire.

[DOCX File , 281 KB - medinform_v10i8e29431_app2.docx]

Multimedia Appendix 3
Ethics approval.
[PDF File (Adobe PDF File), 141 KB - medinform_v10i8e29431_app3.pdf]

References

1. Health Metrics Network, World Health Organization. Framework and standards for country health information systems. 2nd edition. World Health Organization. 2008. URL: https://apps.who.int/iris/handle/10665/43872 [accessed 2020-08-08]
2. Everybody's business -- strengthening health systems to improve health outcomes : WHO's framework for action. World Health Organization. 2007. URL: https://apps.who.int/iris/handle/10665/43918 [accessed 2020-08-08]
3. Crehan C, Kesler E, Nambiar B, Dube Q, Lufesi N, Giaccone M, et al. The NeoTree application: developing an integrated mHealth solution to improve quality of newborn care and survival in a district hospital in Malawi. BMJ Glob Health 2019 Jan 16;4(1):e000860 [FREE Full text] [doi: 10.1136/bmjgh-2018-000860] [Medline: 30713745]
4. Adaletey DL. Leveraging on cloud technology for reporting maternal and child health services at the community level in Ghana. J Health Inf Africa 2018 Jan 21;4(2):12-26 [FREE Full text] [doi: 10.12856/JHIA-2017-v4-i2-146]
5. Agarwal S, Perry HB, Long L, Labrique AB. Evidence on feasibility and effective use of mHealth strategies by frontline health workers in developing countries: systematic review. Trop Med Int Health 2015 Aug;20(8):1003-1014 [FREE Full text] [doi: 10.1111/tmi.12525] [Medline: 25881735]
6. Williams F, Boren SA. The role of the electronic medical record (EMR) in care delivery development in developing countries: a systematic review. Inform Prim Care 2008;16(2):139-145 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2008.01.016] [Medline: 18713530]
7. Tilahun B, Fritz F. Modeling antecedents of electronic medical record system implementation success in low-resource setting hospitals. BMC Med Inform Decis Mak 2015 Aug 01;15:61 [FREE Full text] [doi: 10.1186/s12911-015-0192-0] [Medline: 26231051]
8. Syzdykova A, Malta A, Zolfo M, Diro E, Oliveira JL. Open-source electronic health record systems for low-resource settings: systematic review. JMIR Med Inform 2017 Nov 13;5(4):e44 [FREE Full text] [doi: 10.2196/medinform.8131] [Medline: 29133283]
9. Biruk S, Yilma T, Andualem M, Tilahun B. Health Professionals' readiness to implement electronic medical record system at three hospitals in Ethiopia: a cross sectional study. BMC Med Inform Decis Mak 2014 Dec 12;14:115 [FREE Full text] [doi: 10.1186/s12911-014-0115-5] [Medline: 25495757]
10. Haskew J, Rø G, Saito K, Turner K, Odhiambo G, Wamae A, et al. Implementation of a cloud-based electronic medical record for maternal and child health in rural Kenya. Int J Med Inform 2015 May;84(5):349-354. [doi: 10.1016/j.ijmedinf.2015.01.005] [Medline: 25670229]
11. Ellingsen G, Monteiro E. The organizing vision of integrated health information systems. Health Informatics J 2008 Sep;14(3):223-236 [FREE Full text] [doi: 10.1177/1081180X08093333] [Medline: 18775828]
12. Feroz A, Kadir MM, Saleem S. Health systems readiness for adopting mhealth interventions for addressing non-communicable diseases in low- and middle-income countries: a current debate. Glob Health Action 2018;11(1):1496887 [FREE Full text] [doi: 10.1080/16549716.2018.1496887] [Medline: 30040605]
13. Allen C, Jazayeri D, Miranda J, Biondich PG, Mamlin BW, Wolfe BA, et al. Experience in implementing the OpenMRS medical record system to support HIV treatment in Rwanda. Stud Health Technol Inform 2007;129(Pt 1):382-386. [Medline: 17911744]
14. Zargaran E, Schuurman N, Nicol AJ, Matzopoulos R, Cinnamon J, Taulu T, et al. The electronic Trauma Health Record: design and usability of a novel tablet-based tool for trauma care and injury surveillance in low resource settings. J Am Coll Surg 2014 Jan;218(1):41-50. [doi: 10.1016/j.jamcollsurg.2013.10.001] [Medline: 24355875]
15. Seymour RP, Tang A, DeRiggi J, Munyaburanga C, Cuckovitch R, Nyirishema P, et al. Training software developers for electronic medical records in Rwanda. Stud Health Technol Inform 2010;160(Pt 1):585-589. [Medline: 20841754]
16. Odekunle FF, Odekunle RO, Shankar S. Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region. Int J Health Sci (Qassim) 2017;11(4):59-64 [FREE Full text] [Medline: 29085270]
17. Jawhari B, Keenan L, Zakus D, Ludwick D, Isaac A, Saleh A, et al. Barriers and facilitators to Electronic Medical Record (EMR) use in an urban slum. Int J Med Inform 2016 Oct;94:246-254 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.07.015] [Medline: 27573333]
18. Luna D, Almerares A, Mayan 3rd JC, González Bernaldo de Quirós F, Otero C. Health informatics in developing countries: going beyond pilot practices to sustainable implementations: a review of the current challenges. Healthc Inform Res 2014 Jan;20(1):3-10 [FREE Full text] [doi: 10.4258/hir.2014.20.1.3] [Medline: 24627813]

19. Kruse C, Betancourt J, Ortiz S, Valdes Luna SM, Bamrah IK, Segovia N. Barriers to the use of mobile health in improving health outcomes in developing countries: systematic review. J Med Internet Res 2019 Oct 09;21(10):e13263 [FREE Full text] [doi: 10.2196/13263] [Medline: 31593543]
20. Odendaal WA, Anstey Watkins J, Leon N, Goudge J, Griffiths F, Tomlinson M, et al. Health workers' perceptions and experiences of using mHealth technologies to deliver primary healthcare services: a qualitative evidence synthesis. Cochrane Database Syst Rev 2020 Mar 26;3(3):CD011942 [FREE Full text] [doi: 10.1002/14651858.CD011942.pub2] [Medline: 32216074]
21. Asah F, Kanjo C, Nielsen P. The paradox of technology implementation in health facilities: case of Ghana e-tracker. In: Proceedings of the 3rd International Conference on ICT for African Development. 2019 Presented at: ICT4AD '19; November 26-28, 2019; Yaounde, Cameroon URL: https://www.researchgate.net/publication/339311430_The_Paradox_of_Technology_Implementation_in_Health_Facilities_Case_of_Ghana_e-Tracker
22. Landis-Lewis Z, Manjomo R, Gadabu OJ, Kam M, Simwaka BN, Zickmund SL, et al. Barriers to using eHealth data for clinical performance feedback in Malawi: a case study. Int J Med Inform 2015 Oct;84(10):868-875 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.07.003] [Medline: 26238704]
23. Fenny A, Asante FA, Arhinful DK, Kusi A, Parmar D, Williams G. Who uses outpatient healthcare services under Ghana's health protection scheme and why? BMC Health Serv Res 2016 May 10;16:174 [FREE Full text] [doi: 10.1186/s12913-016-1429-z] [Medline: 27164825]
24. Edum-Fotwe E, Abbey M, Osei I, Hodgson A. Experiences and perceptions of health staff on applying information technology for health data management in Ghana. Measure Evaluation. 2019. URL: https://www.measureevaluation.org/resources/publications/wp-18-224 [accessed 2022-02-10]
25. Nwankwo B, Sambo MN. Can training of health care workers improve data management practice in health management information systems: a case study of primary health care facilities in Kaduna State, Nigeria. Pan Afr Med J 2018 Aug 24;30:289 [FREE Full text] [Medline: 30637073]
26. Awol SM, Birhanu AY, Mekonnen ZA, Gashu KD, Shiferaw AM, Endehabtu BF, et al. Health professionals' readiness and its associated factors to implement electronic medical record system in four selected primary hospitals in Ethiopia. Adv Med Educ Pract 2020 Feb 21;11:147-154 [FREE Full text] [doi: 10.2147/AMEP.S233368] [Medline: 32110135]

Abbreviations

CHN: community health nurse
CHO: community health officer
GHS: Ghana Health Service
HMIS: health management information system
KAP: knowledge, attitude, and practice
LMIC: low- and middle-income country
MCH: maternal and child health
mHealth: mobile health
OR: odds ratio

Edited by C Lovis, J Hefner; submitted 19.04.21; peer-reviewed by E Kesler, D Palazuelos, I Mircheva, M Randriambelonoro; comments to author 02.07.21; revised version received 27.11.21; accepted 25.07.22; published 31.08.22.

Please cite as:
Lee YJ, Lee S, Kim S, Choi W, Jeong Y, Rhim NJJ, Seo I, Kim SY
An mHealth-Based Health Management Information System Among Health Workers in Volta and Eastern Regions of Ghana: Pre-Post Comparison Analysis
JMIR Med Inform 2022;10(8):e29431
URL: https://medinform.jmir.org/2022/8/e29431
doi:10.2196/29431
PMID:36044256

Original Paper

National Development and Regional Differences in eHealth Maturity in Finnish Public Health Care: Survey Study

Jari Haverinen[1,2], MSc, MHSc; Niina Keränen[1,3], MHSc, MD; Timo Tuovinen[1,3], MD; Ronja Ruotanen[1], MHSc; Jarmo Reponen[1,3], MD, PhD

[1]FinnTelemedicum, Research Unit of Medical Imaging, Physics and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland

[2]Finnish Coordinating Center for Health Technology Assessment, Oulu University Hospital, Oulu, Finland

[3]Medical Research Center Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland

Corresponding Author:
Jari Haverinen, MSc, MHSc
FinnTelemedicum, Research Unit of Medical Imaging, Physics and Technology
Faculty of Medicine
University of Oulu
Aapistie 5 A
Oulu, 90220
Finland
Phone: 358 505680144
Email: jari.haverinen@oulu.fi

Abstract

Background: eHealth increasingly affects the delivery of health care around the world and the quest for more efficient health systems. In Finland, the development of eHealth maturity has been systematically studied since 2003, through surveys conducted every 3 years. It has also been monitored in several international studies. The indicators used in these studies examined the availability of the electronic patient record, picture archiving and communication system, health information exchange, and other key eHealth functionalities.

Objective: The first aim is to study the national development in the maturity level of eHealth in primary health care and specialized care between 2011 and 2020 in Finland. The second aim is to clarify the regional differences in the maturity level of eHealth among Finnish hospital districts in 2020.

Methods: Data for this study were collected in 2011, 2014, 2017, and 2020, using web-based questionnaires from the *Use of information and communication technology surveys in Finnish health care* project. In total, 16 indicators were selected to describe the status of eHealth, and they were based on international eHealth studies and Finnish eHealth surveys in 3 areas: applications, regional integration, and data security and information and communications technology skills. The indicators remain the same in all the study years; therefore, the results are comparable.

Results: All the specialized care organizations (21/21, 100%) in 2011, 2014, 2017, and 2020 participated in the study. The response rate among primary health care organizations was 86.3% (139/161) in 2011, 88.2% (135/153) in 2014, 85.8% (121/141) in 2017, and 95.6% (130/136) in 2020. At the national level, the biggest developments in eHealth maturity occurred between 2011 and 2014. The development has since continued, and some indicators have been saturated. Primary health care lags behind specialized care organizations, as measured by all the indicators and throughout the period under review. Regionally, there are differences among different types of organizations.

Conclusions: eHealth maturity has steadily progressed in Finland nationally, and its implementation has also been promoted through various national strategies and legislative changes. Some eHealth indicators have already been saturated and achieved an intensity of use rate of 100%. However, the scope for development remains, especially in primary health care. As Finland has long been a pioneer in the digitalization of health care, the results of this study show that the functionalities of eHealth will be adopted in stages, and deployment will take time; therefore, national eHealth strategies and legislative changes need to be implemented in a timely manner. The comprehensive sample size used in this study allows a regional comparison in the country, compared with previous country-specific international studies.

(JMIR Med Inform 2022;10(8):e35612) doi:10.2196/35612

KEYWORDS

eHealth; electronic health records; picture archiving and communication systems; health information exchange; electronic prescribing; referral and consultation; videoconferencing; clinical decision support systems; health informatics; clinical informatics

Introduction

Background

The World Health Organization (WHO) defines eHealth as "the cost-effective and secure use of information and communications technologies (ICTs) in support of health and health-related fields, including health care services, health surveillance, health literature, and health education, knowledge and research" [1]. According to the WHO, eHealth has a clear and growing impact on the delivery of health care around the world and making health systems more efficient [1]. However, the use of ICTs in health care requires strategic and comprehensive national action to make the best use of it [1,2]. In practice, the term *eHealth* includes a wide range of applications from electronic patient record (EPR) to e–appointment booking (e-booking); therefore, eHealth maturity defines how these different applications were adopted [2-12].

The level of maturity and development of eHealth in health care has been monitored in several international studies [2,3,5-12]. One of the first comprehensive studies was published in 2011. It provided comparative information on the maturity level of eHealth in various European countries [3]. New study reports on eHealth maturity in health care have since been produced by the European Commission (EC), WHO, and Organization for Economic Co-operation and Development (OECD) [2,5,6,8-13]. The Nordic eHealth Research Network has produced comparative information on eHealth maturity levels in various Nordic countries [4,14-16]. In Finland, the development of health care organizations' digitalization has been systematically studied since 2003, through surveys conducted every 3 years [17-21]. The most recent study was conducted in 2021 as part of the *Monitoring and Evaluation of Social and Health Care Information System Services* project [20,22]. It described the situation of digitalization in Finnish health care organizations in 2020 [20]. Although studies have mainly focused on the eHealth maturity level of health care service organizations, eHealth services provided for citizens have also been the subject of research, both in Finland and internationally [19-21,23,24].

Deloitte 2011 report included the EPR, picture archiving and communication system (PACS), e-prescribing, e-referral, e-booking, and telemonitoring (the possibility to use patients' own health data) as the main applications describing the state of health care digitalization [3]. Moreover, the report examined how different countries implemented the wireless use of EPR [3]. EC studies also highlighted the abovementioned applications and the clinical decision support systems (CDSSs) as key indicators of eHealth maturity [5,10]. The degree of integration of CDSS can vary from a separate database to integration with the existing EPR, and Finnish follow-up studies have examined CDSS from this perspective [17-20].

According to a Deloitte report and an EC study, health care integration can be described as an organization's relationship with external service providers, such as other hospitals and health care organizations [3,9]. In the EC study, the sharing of clinical care information, laboratory results, and radiology results between organizations was chosen as the key indicator of integration [5]. They also played an important role in the Finnish health care system, where it is still the case that different organizations largely produce specialized care and primary health care [17-21]. In Finland, Kanta health information exchange (HIE) services are being used from 2010 [25,26]. Although all public health care organizations have now joined Kanta, much of the information exchange continues to use regional HIE (RHIE) systems [18-21,27,28].

Strong user ID is one of the key ways of protecting a patient's health information [10]. Therefore, e-ID and signature were chosen as one of the key indicators for describing data security in Finnish eHealth surveys [17-21]. There must also be sufficient personnel with computer skills to ensure data security practices [17-21]. Technical support for EPR users was chosen as an indicator to describe the reliability of EPR systems [17-21].

General tax revenues collected by the municipalities are the main source of funding for health care and social services in Finland [29]. The state also participates in the costs by paying a general, nonearmarked subsidy to the municipalities [29]. In Finland, municipalities have the primary legal responsibility to organize social and health care services for their residents [29]. Municipalities are responsible for organizing primary health care services for their residents and ensuring that its residents receive the necessary specialized care services [29]. Finland is divided into 21 hospital districts for the provision of specialized care [29]. Every municipality belongs to one of the hospital districts [29]. Decentralized responsibility for organizing health care services has created regional differences in the provision and availability of services [30]. The biggest change in Finnish health care services is the health and social services reform, which will enter into force in 2023 and shift the responsibility for organizing health and social services and rescue services from the municipalities to the 21 new well-being services counties [30]. Some hospital districts have already consolidated their services into a large entity, and in these organizations, specialized care and primary health care fall under the same administrative organization [18-20,31]. The aim of the health and social services reform is to provide equal services to citizens and further develop health care and its operating methods through digitalization [30]. Although the number of EPRs has decreased over the years in both specialized care and primary health care, one of the goals of the health and social services reform is to move toward common solutions for the procurement of EPRs [32,33].

The digitalization of health care has progressed well in Finland [17-21,24-28,33]. This has also come to the fore in international studies, which have highlighted the fact that Finland is one of the pioneers in the digitalization of health care [2,3]. Various national strategies and legal changes have also promoted the

implementation of digitalization in Finnish health care [17-21,25,26]. This study aimed to provide information about eHealth maturity from the perspective of national development and regional differences. The data from this study can be used to examine how eHealth maturity has progressed nationally and in different hospital districts before the health and social services reform [30]. As the digitalization of health care has long been ongoing in Finland, the results can also be exploited internationally. The results show which application areas will be adopted first and how national strategies and legislative changes can contribute to the development of eHealth maturity, both nationally and regionally.

Objectives

The main aims of the study were the following:

1. To study the national development in the maturity level of eHealth in primary health care and specialized care between 2011 and 2020
2. To clarify the regional differences in the maturity level of eHealth among hospital districts in 2020

Methods

Data Collection

This study used data collected in connection with the *Use of information and communication technology in health care 2020* survey and previous surveys in 2011, 2014, and 2017 [17-20]. The data for this study were collected from Finnish public health care providers. The target group for specialized care comprised all 21 hospital districts. In primary health care, the target group included all organizations specified as either independent municipalities or co-operation consortiums of municipalities with the responsibility to provide primary health care services.

The data for the surveys were collected during the first quarters of 2011, 2014, 2017, and 2020, using web-based questionnaires (Webropol; Webropol Ltd). The questions were kept comparable between the survey years. Medical directors and IT leaders (chief information officers) in specialized health care and chief physicians in primary health care were the survey respondents. The questionnaires were sent to them through email. The responses from the entire organizational level were compiled. In some hospital districts, specialized care is also responsible for the municipalities' primary health care services. In these cases, the questionnaire was sent only for specialized care, and the responses regarding specialized care were transferred to the surveys for primary health care.

Table 1 presents the health care organizations that participated in the survey over different years. All specialized care organizations (21/21, 100%) responded to the questionnaire during the survey period. Municipal health care arrangement models changed during the survey years, creating variability in the number of primary health care organizations that participated in the survey and in the response rates and population coverage in different years.

Table 1. Health care organizations participating in the survey in different years.

Year	Respondents in specialized care (n=21), n (%)	Primary health care	
		Respondents, n (%)	Population coverage, %
2011	21 (100)	139 (86.3)[a]	91
2014	21 (100)	135 (88.2)[b]	95
2017	21 (100)	121 (85.8)[c]	95
2020	21 (100)	130 (95.6)[d]	99

[a]Sample size, n=161.

[b]Sample size, n=153.

[c]Sample size, n=141.

[d]Sample size, n=136.

Indicators for eHealth and Their Analysis

In total, 16 indicators were selected to describe the status of eHealth (Table 2). They were based on the indicators in the eHealth report on specialized care, EC eHealth studies, and Finnish eHealth surveys in three areas: (1) applications, (2) regional integration, and (3) data security and ICT skills (Table 2) [3,5,10,17-21]. Traditionally, eHealth surveys have used the availability of applications or services as indicators [3]. However, availability saturation has been achieved in Finland in several health care application areas. For example, in 2010, EPR was available in all specialized care and primary health care organizations [17]. For several years, Finnish national eHealth surveys have also enquired about the intensity of use to describe the integration of an application or service into normal health care operations [17-21]. More specifically, it describes which proposition of a specific service is provided through eHealth means within an organization. For the intensity of use, the percentages (0%, 25%, 50%, 90%, 99%, and 100%) were chosen to correspond to the verbal answers, "not in use," "a quarter," "half," "most," "almost all," and "all," respectively. In the summary of indicators, the mean value of the results of the participating organizations was displayed. Where possible, the intensity of use of application was selected as an indicator to describe the use of eHealth. This better describes the deployment of eHealth in situations in which the functionality is already widely available.

Table 2. Indicators for eHealth maturity.

Areas and functionalities	Indicator	Responses
Applications		
EPR[a]	Intensity of use	0=not in use, 2=≤25%, 4=≤50%, 7=≤90%, 9.9=≤99%, and 10=100%
Wireless use of EPR	Availability (local or external)	0=not available and 10=available
Picture archiving and communication system	Intensity of use	0=not in use, 2=≤25%, 4=≤50%, 7=≤90%, 9.9=≤99%, and 10=100%
Clinical decision support system	Integration level—average between the integration of diagnostic support and a drug interaction system	0=not available, 4=stand-alone web-based database on desktop, 6=database with access by navigating from the EPR, 8=automatic displayer of selected items, and 10=system for automatic integration of the EPR and database
e-Prescribing	Availability	0=not available and 10=available
e-Referral	Intensity of use	0=not in use, 2=≤25%, 4=≤50%, 7=≤90%, 9.9=≤99%, and 10=100%
Consultation e-referral	Intensity of use	0=not in use, 2=≤25%, 4=≤50%, 7=≤90%, 9.9=≤99%, and 10=100%
Teleconsultations via videoconferencing	Intensity of use—how often has the service been in use?	0=not in use, 4=less often, and 10=during the past 3 months
Possibility to use patients' own health data	Availability	0=not available and 10=available
e–Appointment booking	Intensity of use—the patient selects an appointment time on their terminal (eg, computer) and it is transferred directly to the system	0=0% to 10=100%
Regional integration		
Exchange of clinical care information[b]	Availability	0=not available and 10=available
Exchange of laboratory results[b]	Availability	0=not available and 10=available
Exchange of radiology reports[b]	Availability	0=not available and 10=available
Information security and ICT[c] skills		
e-ID and signature	Availability	0=not available and 10=available
Personnel with computer skills	Proportion	0=0% to 10=100%
Technical support for EPR	Intensity	0=not in use; 2=occasionally; 5=daily, but for less than normal office hours; 7=during normal office hours; and 10=at all times during the opening hours of the organization

[a]EPR: electronic patient record.

[b]Health information exchange outside the centralized national Kanta services.

[c]ICT: information and communications technology.

Ethical Consideration

The study followed the guidelines of the Finnish Advisory Board on Research Integrity [34]. The respondents were informed about the study, and they answered as representatives of the organizations being studied. No sensitive personal information was collected. The data were processed and stored in a secure environment, according to the procedures of the University of Oulu.

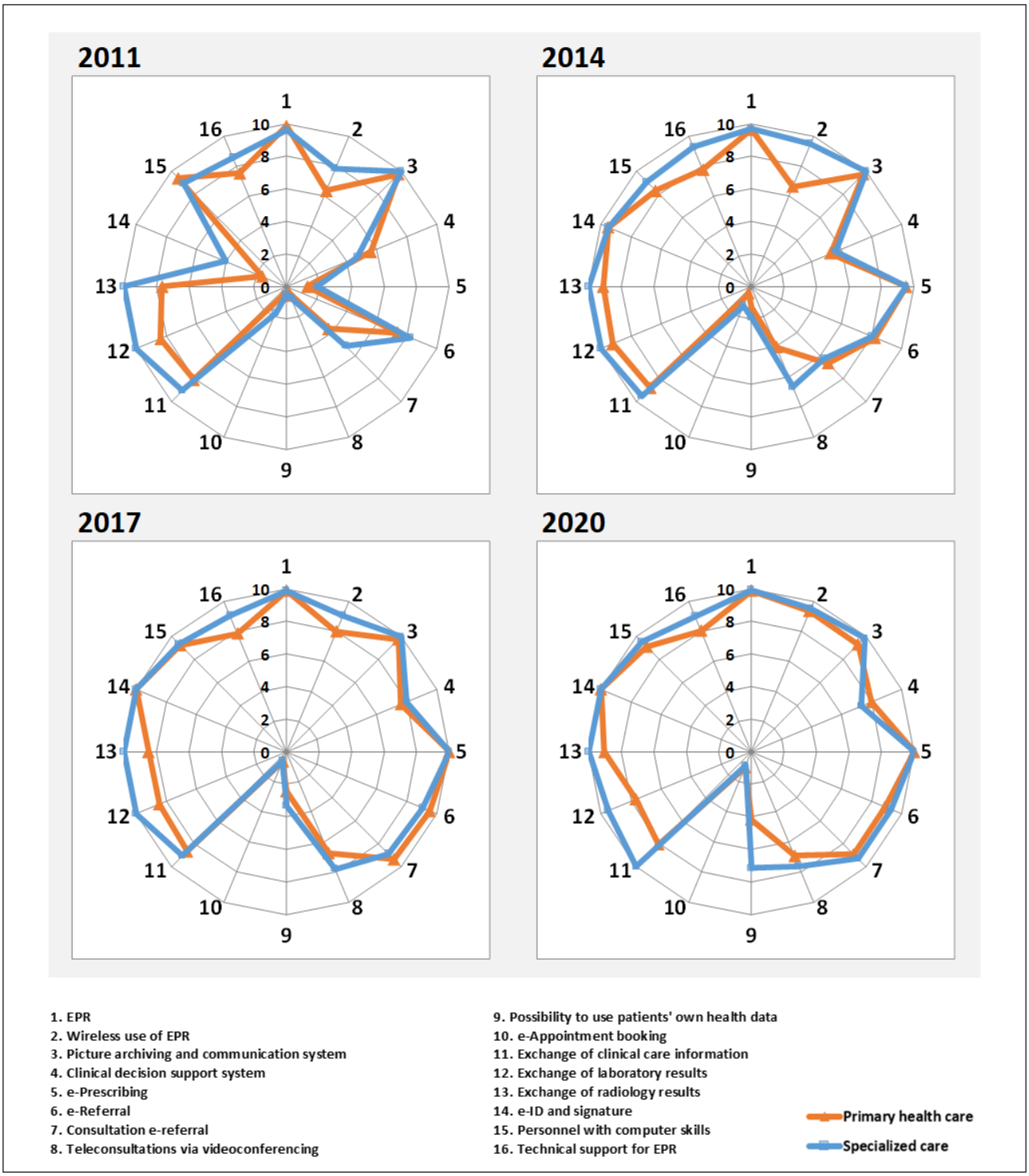
Results

Maturity of eHealth in Specialized Care and Primary Health Care Organizations at the National Level

Overview

Figure 1 presents the national development in eHealth maturity in specialized care and primary health care organizations between 2010 and 2020. The results show that primary health care is generally behind specialized care organizations, as measured by all indicators and throughout the period under review. The biggest difference can be seen in the area of RHIE.

Figure 1. The national development in the maturity level of eHealth in the years 2011, 2014, 2017, and 2020 (modified from the studies by Reponen et al [19,20]). EPR: electronic patient record.



Applications

The EPR's intensity of use has been at a high level since 2011, in both specialized care and primary health care organizations. There has been no significant change over the years. The wireless use of EPR has been available since 2014 in approximately all specialized care organizations (21/21, 100%), and there has been no significant change by 2020. In primary health care, availability has grown steadily, and in 2020, it has reached the same level as in specialized care. The intensity of use of PACS has been high in both specialized care organizations and primary health care centers in Finland since 2011. The integration level of the CDSS has increased since 2011; however, no growth can be seen after 2017. The level of integration has remained the same in both specialized care and primary health care organizations throughout the period under review.

The most significant change in the availability of e-prescribing occurred between 2011 and 2014. Since 2014, e-prescribing has been widely available in both specialized care organizations and primary health care centers. There was no significant change

in the intensity of use of e-referral between 2011 and 2014. Since 2017, the intensity of use of e-referral has been high in both specialized care and primary health care organizations. The intensity of use of consultation e-referral increased between 2011 and 2017 in both specialized care and primary health care organizations. No significant change was seen by 2020.

In 2011, the intensity of use of teleconsultations via videoconferencing was very low, but significant growth was observed in both specialized care and primary health care organizations in the 2014 survey. In 2017, the intensity of use increased slightly, but remained at the same level in the 2020 survey. The possibility to use the patient's own health data remained very limited in 2011 and 2014. There has since been significant growth in this area of application, especially in specialized care organizations. The intensity of use of e-booking remains low in the 2020 survey, and no growth can be seen throughout the survey period.

Regional Integration

All specialized care organizations reported that exchange of laboratory results and radiology reports was available from 2011. However, in 2020, not all specialized care organizations reported that regional information exchange of laboratory results was available. Exchange of clinical care information was unavailable in all specialized care organizations during the survey period, except in the 2020 survey, when all organizations (21/21, 100%) reported that it was available.

In 2020, approximately 80.1% (109/136) of the primary health care centers reported the availability of regional information exchange in all 3 areas of information exchange. There is no major change in the results of the 2020 survey compared with those of the 2010 survey. The highest reported availability of regional information exchange in primary health care centers was observed in the 2014 survey.

Data Security and ICT Skills

The availability of e-ID and signature was low in 2010, in both specialized care organizations and primary health care centers. In 2014, it was available in approximately all specialized care organizations (20/21, 95%) and primary health care centers (146/153, 95.4%), and it has been in use in all specialized care organizations (21/21, 100%) and primary health care centers (141/141, 100%) since 2017. Throughout the survey period, organizations have reported that the number of personnel with computer skills was approximately 90%. There are slight variations in the reported results among different years. EPR technical support at all times during the organization's opening hours remains unavailable in some specialized care organizations (7/21, 33%) in 2020. No significant change can be seen in the results during the survey period.

eHealth Profiles at the Regional Level

The status of the eHealth profiles of different types of health care organizations is presented in Figure 2.

Especially in primary health care, the results do not show that the eHealth maturity level is better in large university hospital districts than in smaller hospital districts (Figure 2). For example, in university hospital districts, the availability of EPR technical support and wireless use of EPR is lower than the average of the other hospital districts. The availability of RHIE is also low in university hospital districts, especially for primary health care. Only the intensity of use of teleconsultations via videoconferencing is at a higher level than that in the university hospital districts.

In 43% (9/21) of the hospital districts, primary health care and specialized care are under the same administrative organization, and all these organizations use the same EPR brand throughout their municipalities and specialized care organizations (Figure 2). Compared with the other hospital districts, the combined organizations report better results for the availability of the wireless use of EPR and RHIE and the intensity of use of e-referral and consultation e-referral. The use rates of e-referral and consultation e-referral are low only in Kainuu. South and North Karelia still stand out from these organizations because of their good results. In both hospital districts, all indicators are saturated, except for the use of e-booking. In North Karelia, there is also scope for improvement in the number of personnel with computer skills.

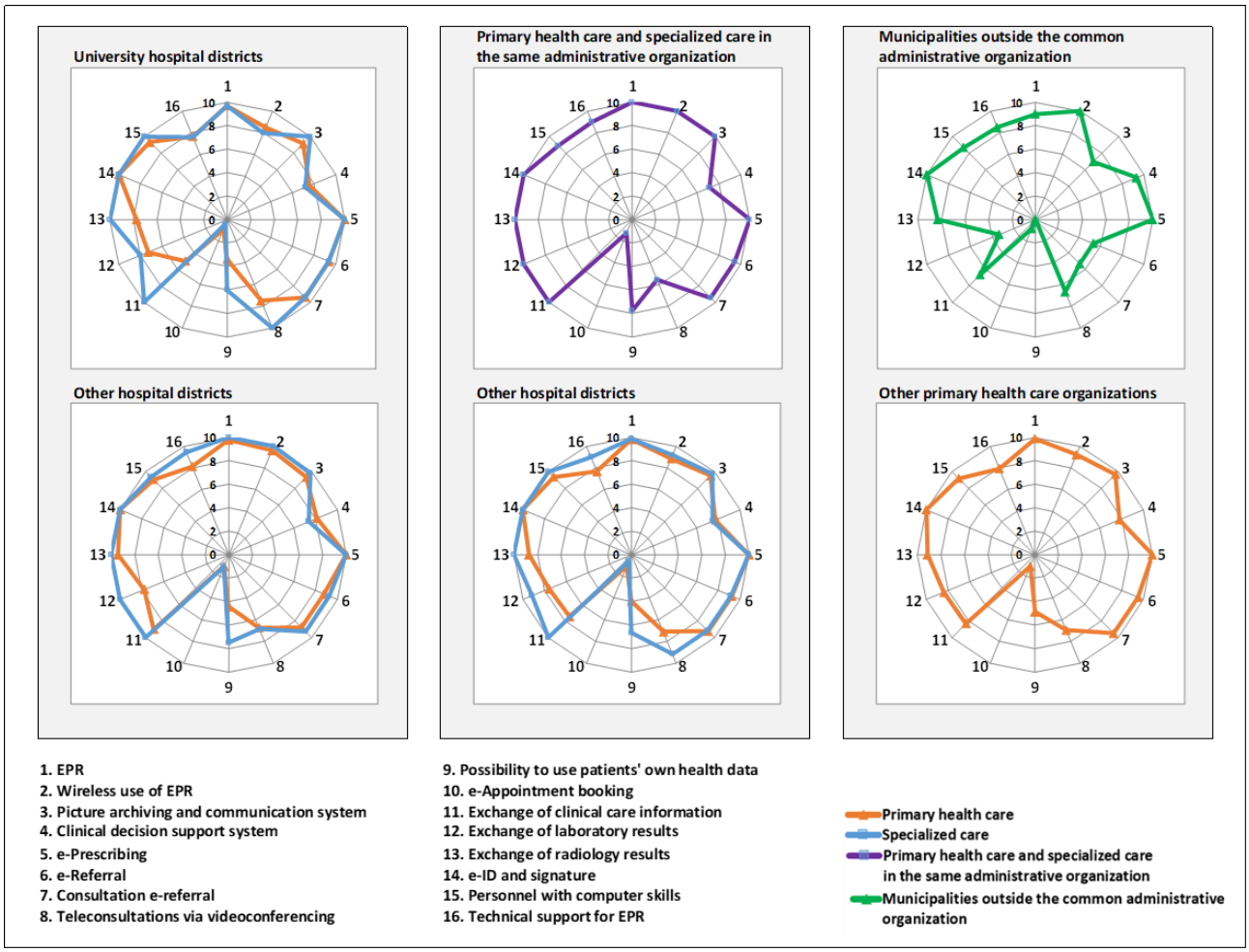
In total, 14% (3/21) of the hospital districts also have individual municipalities outside the common administrative organization (Figure 2). In this 14% (3/21) cases, the results obtained from the municipalities outside the common organizations are worse than those at the national level. The results are particularly worse in the intensity of use of e-referral and consultation e-referral and the availability of laboratory result exchange. However, the integration level of the CDSS is higher in these municipalities than in the other primary health care organizations.

In addition to the 43% (9/21) of the hospital districts that have primary and secondary care under the same organization, 10% (2/21) of the districts reported that they were using the same EPR brand in both specialized care and primary health care organizations in their area, making a total of 52% (11/21). In all these organizations, the use rate of EPR was 100% and the wireless use of EPR was available. They reported better results in RHIE, especially when comparing with primary health care results. All these hospital districts, except Kainuu and Kanta-Häme, reported 100% use rates for e-referral and consultation e-referral. Of the 11 organizations, 8 (73%) organizations reported that EPR technical support was always available during their organization's opening hours. Of the 10 hospital districts without the same EPR in use in their area, 6 (60%) districts reported that EPR technical support was always available during opening hours in specialized care organizations. Therefore, there was no significant difference in the results for specialized care organizations; however, in primary health care, the situation was different. None of the primary health care organizations without the unified EPR (104/136, 76.5%) reported that this service was always available during the organization's opening hours.

The integration level of the CDSS varies greatly among hospital districts. In total, 14% (3/21) of the specialized care organizations reported that the CDSS was a stand-alone web-based database on the computer desktop, and 24% (5/21) reported that they had automatic integration of the EPR. Although the use of patients' own health data significantly

increased by 2020, regional differences in its availability still remained. Overall, 29% (6/21) of the specialized care organizations reported that this application was unavailable in 2020. The regional difference can also be seen in the intensity of use of teleconsultations via videoconferencing, because 19% (4/21) of the specialized care organizations reported that this application was not in use and the remaining 81% (17/21) reported that this service had been in use in the past 3 months. The intensity of use of e-booking remained low throughout the survey period, and no major regional differences can be seen in the 2020 survey results.

Figure 2. The status of the eHealth profiles of different types of health care organizations. EPR: electronic patient record.



Discussion

Principal Findings

The digitalization of health care has progressed well in Finland, and its implementation has also been promoted through various strategies and legislative changes [17-21,25,26]. The progress of health care digitalization has also been systematically monitored through studies since 2003 [17-21]. The studies' timing has aimed for alignment with the schedules of key legislative changes and strategies [17-21]. This survey study presented the development of eHealth maturity, measured by key indicators, in Finnish health care in 2011, 2014, 2017, and 2020. It also studied the regional differences in the maturity level of eHealth among hospital districts in 2020, measured by the same indicators. The study covered all Finnish specialized care organizations and a comprehensive portion of primary care organizations. In every study year, the response rate of primary care organizations was >86% and population coverage was >91%. The comprehensive sample size of this study's respondents allowed regional comparison among organizations. Previous international studies have been based on sample data, so the results are presented country by country and not regionally within a country. For example, in the latest EC benchmarking study on the deployment of eHealth among general practitioners (GPs) in 2018, the sample size of Finland was 2.5% of GPs [8]. Internationally identified eHealth indicators were used in this study, and they remained the same between different study years; thus, the results were comparable. As the digitalization of health care in Finland has progressed, availability saturation has been achieved in some health care application areas [17-21]. In these cases, the intensity of use of the application was selected to describe the use of eHealth instead of availability [17-21].

Nationally, eHealth maturity has progressed steadily in Finland, with the biggest developments in eHealth maturity occurring between 2011 and 2014. On the basis of the national results, the functionalities have been built step by step. The first phase focused on key functionalities such as the deployment of EPR, PACS, and RHIE. Since 2011, the reported intensity of use of EPR has been high, and in the 2020 survey, only 5% (1/21) of

the hospital districts reported the intensity of use of EPR as <100%. The intensity of use of PACS has been 100% in all hospital districts (21/21, 100%) since 2017. The next development step was to focus on the deployment of e-referral and consultation e-referral functionalities. Between 2011 and 2017, the use rate of these functionalities changed greatly, but the development has since stalled.

The uptake of the Finnish Kanta services started in 2010 [25]. Their introduction enabled the use of e-prescribing, which has since been widely adopted [26]. It became mandatory in 2017, which can also be seen in the result—since 2017, all specialized care organizations (21/21, 100%) and primary health care centers (141/141, 100%) have adopted it. The e-ID and signature also became mandatory with the introduction of the Kanta services, and since 2017, they have been introduced in all specialized care organizations (21/21, 100%) and primary health care centers (141/141, 100%). This emphasizes that changes in the law may lead to significant changes in eHealth maturity, as shown by a comparison of the results from 2011 to 2017 [25,26].

The development of smart devices and telecommunication networks has enabled the provision of an increasing number of remote and wireless services [3,19-21]. This is also reflected in the results because the availability of the wireless use of EPR and the use rate of teleconsultations via videoconferencing increased during the survey. The biggest development in these indicators can be seen between 2011 and 2014. The intensity of use of teleconsultations via videoconferencing has reached 80% since 2017, but no major improvement has been observable since then. Regionally, there are differences in the frequency at which this functionality has been used.

The aim of the health and social services reform is to improve the quality of services and ensure regional equality [30]. According to this study, there are still regional differences in the eHealth maturity levels among different hospital districts. There are differences in how comprehensive RHIE can be provided, how well EPR technical support is organized, and whether e-referral and consultation e-referral are widely used. These are particularly evident in the indicators for primary health care, especially in the municipalities outside the common administrative organization. Their results are noticeably worse in the intensity of use of e-referral, consultation e-referral, and RHIE. Regional variation also exists among hospital districts in how well CDSS is integrated into the EPR, the use rate of teleconsultations via videoconferencing, and the use of patients' own health data. However, EPR and PACS are widely used in all specialized care organizations (21/21, 100%), and wireless use of EPR is available in most hospital districts (20/21, 95%). The results indicate that operating under the same administrative organization and using the same EPR brand in the region will enable the support of a more comprehensive level of eHealth maturity regionally for both specialized care and primary health care. On the basis of these results, it seems the goal of the health and social services reform to establish large operational units will improve the opportunities to provide a better and equal level of eHealth maturity [30]. The goal of the health and social services reform to move toward common solutions for the procurement of EPRs seems to help achieve better results in the intensity of use of EPR, availability of wireless use of the EPR, and availability of RHIE [32,33]. According to this study, a unified EPR brand also seems to allow slightly better EPR technical support.

Certain similarities can be observed if we examine the results in terms of how different hospital districts provide eHealth services for their citizens. This is especially true in the cases of South and North Karelia. Ruotanen et al [24] studied the availability of eHealth services for citizens in 2020. According to their study, these 2 hospital districts offer the widest range of eHealth services for citizens and they have the best eHealth maturity [24]. Could the explanation be that determined work has been done in these hospital districts to promote the digitalization of health care and implement national strategies? This may also be because organizations have had time to implement health care integration over a sufficiently long period because, for example, Eksote, the common administrative organization in South Karelia, has been operating since 2010 [31]. This highlights that an early investment in development is important, because moving functionalities into the production phase is a time-consuming process [7,25].

When this study's results are examined internationally, we see that in the surveys conducted by the EC in 2013 and the WHO in 2016, Finland was ahead of the European Union (EU) and global average in the selected indicators [5,6]. For example, during those years, the availability of EPR and PACS was higher than the EU and global average, and even currently, the intensity of use rates for both applications in Finland are approximately 100% [5,6]. The biggest increase has been observed in the use of the patient's own health data, because in the EC study in 2014, the use of the patient's own health data was very limited at the EU level, as in Finland [5]. Moreover, in the Nordic study in 2014, the use of this functionality was low in all the Nordic countries. A Nordic benchmarking study noted that the Nordic countries were eHealth pioneers, especially in the HIE and EPR functionalities [4,5]. The Nordic eHealth Research Network also states in its study that several eHealth functionalities have already reached 100% availability in the Nordic countries; therefore, studies should focus more on the intensity of use of these functionalities [16]. Our study provides an example of how intensity of use data can be collected in a situation in which data on availability alone reveal insufficient details.

The latest benchmarking study results have shown that Finland remains as one of the pioneers in the development of eHealth. Ammenwerth et al [7] performed an international comparative study of 6 basic eHealth indicators across 14 countries in 2020. On the basis of their findings, Finland showed the best overall outcome in all the selected eHealth indicators in the study, followed by South Korea, Japan, and Sweden [7]. According to the study, Finnish health care professionals could easily access their patients' health data and were able to add the data to electronic health records, but the possibility for patients to add data to their health records remains to be improved in Finland [7]. The 2017 OECD eHealth indicator survey, conducted in 38 countries, found that no country outperformed all countries in all the indicators used in the survey, but in contrast, no country lagged behind the other countries, as measured by all the indicators [35]. According to this study, Finland is one of the top performers in the availability of EPR

and use of HIE of radiology results and images [35]. Finland was also noted as a top performer among OECD countries in technical and operational readiness to provide national health information from EPRs [12]. The availability to electronically request prescription renewal or refill and patients' ability to access test results via the web was <50% in approximately all the OECD countries that participated in the 2017 study [35]. The availability of e-booking clearly needs to be improved among OECD countries, because in approximately all countries, including Finland, its availability was <50% [35]. The latest EC eHealth benchmarking study in all EU countries in 2018 highlighted Finland as one of the top performers, especially in the sharing of radiology test images and reports [8]. The EPR has been fully available across all EU countries since 2018. HIE is also well adopted in EU countries, but it has been less adopted than the EPR [8]. There was an increase in the adoption of HIE across all member states between 2013 and 2018, and along with Denmark, Estonia, and Sweden, Finland is among the top clinical data performers in HIE [8]. The highest HIE availability rates among EU countries were reported for receiving laboratory reports (77%), certifying sick leave (69%), sending and receiving referral and discharge letters (53%), and transferring prescriptions to pharmacists (52%) [8]. Compared with these results, this study shows that Finland is ahead of the EU average in the exchange of laboratory results, e-referral, and e-prescribing. Although a study conducted in 2018 found that e-prescribing was widely adopted in the 23 EU countries studied, there was great variation in authentication procedures among the countries [36]. One of the goals of EU for the development of eHealth has been to promote cross-border health care [37]. However, only Finland and Croatia have e-prescribing systems that can prescribe medications to be dispensed abroad [36].

Scope for development remains among EU countries, especially in the adoption of telehealth services and personal health records [8]. There is also scope for improvement among OECD countries in the adoption of telehealth services, because only approximately one-third of the hospitals indicated that they had telehealth capabilities for patient consultation [35]. In any case, consultation with other professionals using telehealth services is well adopted in Finland, because this study indicates that remote consultation via videoconferencing has been extensively adopted. However, there is still scope for improvement in Finland; for example, the use of CDSS was below the EU average in 2018 [8]. e-Booking in the Finnish public health care context clearly needs to be developed. On the basis of this study, no significant development has been seen during the entire 10-year follow-up period. However, 43% of EU GPs reported that their ICT systems allowed their patients to request appointments in 2018; therefore, Finland clearly has scope for improvement in this area [8].

This study was conducted in the Finnish health care environment, but we believe the findings are applicable to other countries that aim to develop health care further through digitalization. On the basis of the results, the deployment of eHealth applications will take time, and both legislative changes and national strategies may help to promote implementation [38]. According to the WHO, 58% of the countries that responded to their global survey in 2016 reported having an eHealth strategy [2]. In Finland, the first national strategy for applying ICT to health care and social welfare was introduced in 1995 by the Ministry of Social Affairs and Health [21]. Thus, Finland was one of the first countries, along with San Marino, Norway, and Canada, to have eHealth strategies or policies in place [2]. Strategies have since been used to promote the structured recording of patient data and the integration between systems and to increase the electronic exchange of information between patients and health care professionals [17-21,25,26]. Payne et al [39] studied the status of HIE among 6 countries, and they stated that the complexity of health care systems will present barriers to HIE. This is the case in Finland, because there are still regions where different organizations provide specialized care and primary health care and use the different EPR brands in their region. The study also noted that in countries that have successfully achieved HIE, the impetus came from the government [39]. In Finland, HIE between organizations has been promoted through the national Kanta services, in which all public health care organizations have joined [25,26]. Despite this national service uptake, which allows information exchange between organizations, there is still a possibility to use RHIE systems, as highlighted in this study [28]. The aim of the latest strategy is also to promote interoperable and modular architectures and information security and to ensure sufficient data connections [40]. Legal changes may also contribute significant improvements to eHealth maturity, as can be seen in this study's results regarding the availability of electronic prescription and e-ID and signature. These functions became mandatory for all public health care organizations in 2017, and the results show significant development between 2011 and 2017 [21,25,26]. However, the implementation of the new functionalities will take time because the path of the Finnish national electronic prescription system from legislation to full implementation took 10 years [25]. A very important step forward to enable RHIE in hospital districts was a law that came into force in 2011, which allowed public health care to build common patient registries for hospital districts and primary health care organizations in each of the regions. After the law's implementation, specific consent from a patient who is informed was no longer required for information retrieval [21].

The results reveal that a basic infrastructure such as the EPR must be in place to enable other advanced functionalities such as the CDSS and HIE, because the structured data storage of EPR is a prerequisite for the operation of CDSS systems [5,10,17-21]. Presumably, an EPR and broadband wireless infrastructure must be available for the wireless use of EPR [17-21]. The results also show that operating under a common regional administrative organization and using the same EPR brand will enable better overall eHealth maturity results, especially in RHIE, for both specialized care and primary health care, at least in the Finnish health care context. Although national strategies can guide the development of eHealth, the regions' own determined work can also lead to even better results. The results highlight a few regions with high degree of eHealth maturity in the selected indicators in this study while providing comprehensive eHealth services to their citizens, as shown in the study by Ruotanen et al [24]. The organization's own activities also affect the extent to which EPR technical

support is provided and whether personnel's ICT skills are promoted through training [2].

Limitations

The results show that not all indicators may be relevant when examining the eHealth maturity of future public health care in Finland. When these eHealth maturity level studies started in Finland, most of the functionalities collected according to internationally used availability indicators were still in the development phase. However, some indicators, such as e-ID and signature and e-prescribing, have been saturated since 2017 and provided no additional information. Therefore, instead of using e-ID, a better indicator for the regional evaluation of data security could be the availability of a documented data security policy or data security plan.

Regarding primary health care, the number of survey respondents has decreased over time. This is explained by the merging of municipalities into large administrative entities. In contrast, the response rate of primary health care centers to the survey has increased during the survey's implementation; thus, the sample size has differed slightly in the different survey years. This may cause minor variations in the results for different years.

The results of this study are based on the data provided by various organizations. In each organization, its management has compiled organization-specific responses from experts in different areas. Different experts may have responded to the survey in different years of the study; therefore, the questions may have been understood differently. However, efforts were made to assist the respondents by providing them with their responses from the previous survey year as a reference. The respondents may have represented the administrative organization; therefore, they may not have had a complete picture of the situation in practice. For example, the interpretation of the proportion of personnel with computer skills may vary among respondents. The interpretations of terms in various years may also vary, depending on what was topical at the time. The intensity of use of certain eHealth applications is based on respondents' estimates rather than log data, meaning that there may be variation in results, depending on the respondent's interpretation.

Conclusions

eHealth maturity has steadily progressed nationally in Finland, and various national strategies and legislative changes have promoted its deployment. The biggest developments in eHealth maturity occurred between 2011 and 2014. Some indicators reached saturation and an intensity of use rate of 100%. However, the scope for development remains, especially in primary health care. Regionally, differences remain among different organizations. Some hospital districts have already been operating under a common administrative organization for a long time, and the results suggest that they will be more prepared for the approaching health and social services reform. The national eHealth strategies and legislative changes need to be implemented in a timely manner, because the results of this study show that the functionalities of eHealth will be adopted in stages and deployment will take time.

Acknowledgments

The study is part of the national *Monitoring and Evaluation of Social and Health Care Information System Services* research project, which is cofunded by the Ministry of Social Affairs and Health (Sosiaali-ja terveysministeriö), and the participants are coordinated by the National Institute for Health and Welfare (Terveyden ja hyvinvoinnin laitos).

Authors' Contributions

All authors participated sufficiently in the study to take public responsibility for appropriate portions of the content. JH, TT, NK, and JR were responsible for the study conception and design. NK and RR performed data acquisition. JH, TT, NK, and JR analyzed and interpreted the data. JH and JR drafted the manuscript.

Conflicts of Interest

NK, RR, and JR have received funding to complete this study as part of the national *Monitoring and Evaluation of Social and Health Care Information System Services* project.

References

1. eHealth. World Health Organization. 2022. URL: http://www.emro.who.int/health-topics/ehealth/ [accessed 2022-05-14]
2. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. Geneva, Switzerland: World Health Organization; 2016. URL: https://apps.who.int/iris/handle/10665/252529 [accessed 2021-12-09]
3. eHealth Benchmarking III, SMART 2009/0022. Deloitte & Ipsos. 2011 Apr 13. URL: https://tinyurl.com/3n4md9d8 [accessed 2021-12-07]
4. Hyppönen H, Kangas M, Reponen J, Nøhr C. Nordic eHealth Benchmarking. Nordic Council of Ministers, Nordic Council of Ministers Secretariat. 2015. URL: http://norden.diva-portal.org/smash/record.jsf?pid=diva2%3A821230&dswid=-444 [accessed 2021-12-08]
5. European hospital survey: benchmarking deployment of e-health services (2012–2013). Final report. Joint Research Centre, Institute for Prospective Technological Studies. 2014. URL: https://data.europa.eu/doi/10.2791/56790 [accessed 2021-11-30]

6. Atlas of eHealth country profiles: the use of eHealth in support of universal health coverage – Based on the findings of the third global survey on eHealth 2015. World Health Organization. 2016 Jan 1. URL: https://www.who.int/publications/i/item/9789241565219 [accessed 2021-12-07]
7. Ammenwerth E, Duftschmid G, Al-Hamdan Z, Bawadi H, Cheung NT, Cho K, et al. International comparison of six basic eHealth indicators across 14 countries: an eHealth benchmarking study. Methods Inf Med 2020 Dec;59(S 02):e46-e63 [FREE Full text] [doi: 10.1055/s-0040-1715796] [Medline: 33207386]
8. European Commission, Directorate-General for Communications Networks, Content and Technology, Folkvord F, Hocking L, Altenhofer M. Benchmarking deployment of eHealth among general practitioners (2018) : final report. Publications Office of the European Union. 2019. URL: https://op.europa.eu/en/publication-detail/-/publication/d1286ce7-5c05-11e9-9c52-01aa75ed71a1 [accessed 2021-12-09]
9. Joint Research Centre, Institute for Prospective Technological Studies, Sabes-Figuera R. European hospital survey : benchmarking deployment of e-health services (2012–2013) : country reports. Publications Office of the European Union. 2014. URL: https://data.europa.eu/doi/10.2791/55973 [accessed 2021-12-08]
10. Joint Research Centre, Institute for Prospective Technological Studies, Sabes-Figuera R, Maghiros I. European hospital survey : benchmarking deployment of e-health services (2012–2013) : composite indicators on eHealth deployment and on availability and use of eHealth functionalities. Publications Office of the European Union. 2014. URL: https://data.europa.eu/doi/10.2791/56511 [accessed 2021-12-06]
11. Oliveira Hashiguchi T. Bringing health care to the patient: an overview of the use of telemedicine in OECD countries. OECD Health Working Papers. 2020. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/bringing-health-care-to-the-patient_8e56ede7-en [accessed 2021-12-07]
12. Oderkirk J. Readiness of electronic health record systems to contribute to national health information and research. OECD Health Working Papers. 2017. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/readiness-of-electronic-health-record-systems-to-contribute-to-national-health-information-and-research_9e296bf3-en [accessed 2021-11-28]
13. Draft OECD guide to measuring ICTs in the health sector. Organization for Economic Co-operation and Development (OECD). 2015. URL: https://www.oecd.org/health/health-systems/Draft-oecd-guide-to-measuring-icts-in-the-health-sector.pdf [accessed 2021-12-09]
14. Hyppönen H, Faxvaag A, Gilstad H, Hardardottir GA, Jerlvall L, Kangas M, et al. Nordic eHealth indicators: organisation of research, first results and plan for the future. Stud Health Technol Inform 2013;192:273-277. [Medline: 23920559]
15. Hyppönen H. Nordic eHealth benchmarking: from piloting towards established practice. TemaNord, Nordic Council of Ministers. 2017. URL: https://www.oecd-ilibrary.org/content/publication/tn2017-528 [accessed 2021-12-06]
16. Nøhr C, Koch S, Vimarlund V, Gilstad H, Faxvaag A, Hardardottir GA, et al. Monitoring and benchmarking eHealth in the Nordic countries. Stud Health Technol Inform 2018;247:86-90. [Medline: 29677928]
17. Winblad IRJ, Reponen J, Hämäläinen P. Tieto- ja viestintäteknologian käyttö terveydenhuollossa vuonna 2011 : Tilanne ja kehityksen suunta. Terveyden ja hyvinvoinnin laitos (THL). 2012. URL: http://urn.fi/URN:NBN:fi-fe201205085463 [accessed 2021-12-09]
18. Reponen J, Kangas M, Hämäläinen P, Keränen N. Tieto- ja viestintäteknologian käyttö terveydenhuollossa vuonna 2014 - Tilanne ja kehityksen suunta. Terveyden ja hyvinvoinnin laitos (THL). 2015. URL: http://urn.fi/URN:ISBN:978-952-302-486-1 [accessed 2021-12-09]
19. Reponen J, Kangas M, Hämäläinen P, Keränen N, Haverinen J. Tieto- ja viestintäteknologian käyttö terveydenhuollossa vuonna 2017 : Tilanne ja kehityksen suunta. Terveyden ja hyvinvoinnin laitos (THL). 2018. URL: http://urn.fi/URN:ISBN:978-952-343-108-9 [accessed 2021-12-09]
20. Reponen J, Keränen N, Ruotanen R, Tuovinen T, Haverinen J, Kangas M. Tieto- ja viestintäteknologian käyttö terveydenhuollossa vuonna 2020 : Tilanne ja kehityksen suunta. Terveyden ja hyvinvoinnin laitos (THL). 2021. URL: http://urn.fi/URN:ISBN:978-952-343-771-5 [accessed 2021-06-02]
21. Vehko T, Ruotsalainen S, Hyppönen H. E-health and e-welfare of Finland : Check Point 2018. National Institute for Health and Welfare. 2019. URL: https://urn.fi/URN:ISBN:978-952-343-326-7 [accessed 2021-06-02]
22. Sosiaali- ja terveydenhuollon tietojärjestelmäpalveluiden seuranta ja arviointi (STePS 3.0). Terveyden ja hyvinvoinnin laitos. 2020. URL: https://thl.fi/fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/sosiaali-ja-terveydenhuollon-tietojarjestelmapalveluiden-seuranta-ja-arviointi-steps-3.0- [accessed 2021-12-09]
23. From innovation to implementation: eHealth in the WHO European Region. World Health Organization. 2016. URL: https://apps.who.int/iris/handle/10665/326317 [accessed 2021-07-07]
24. Ruotanen R, Kangas M, Tuovinen T, Keränen N, Haverinen J, Reponen J. Finnish e-health services intended for citizens – national and regional development. Finnish J EHealth EWelfare 2021 Oct 25;13(3):283-301. [doi: 10.23996/fjhw.109778]
25. Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010–2017: a prospective, longitudinal, indicator-based study. Finnish J EHealth EWelfare 2018 Dec 04;10(4):381-395. [doi: 10.23996/fjhw.74511]
26. Jormanainen V. Implementation and adoption of the national Kanta services in community pharmacies and municipality public health care centres in 2010–2016 in Finland. Finnish J EHealth EWelfare 2019 May 05;11(3):169-182. [doi: 10.23996/fjhw.77601]

27. Hyppönen H, Lumme S, Reponen J, Vänskä J, Kaipio J, Heponiemi T, et al. Health information exchange in Finland: usage of different access types and predictors of paper use. Int J Med Inform 2019 Feb;122:1-6. [doi: 10.1016/j.ijmedinf.2018.11.005] [Medline: 30623778]
28. Keränen N, Tuovinen T, Haverinen J, Ruotanen R, Reponen J. Regional health information exchange outside of the centralized national services for public health care in Finland: a national survey. Finnish J EHealth EWelfare 2022 Apr 14;14(1):31-42. [doi: 10.23996/fjhw.111775]
29. Social and health services. Ministry of Social Affairs and Health. 2022. URL: https://stm.fi/en/social-and-health-services [accessed 2022-05-14]
30. What is the health and social services reform. Valtioneuvosto | Finnish Government. 2021. URL: https://soteuudistus.fi/en/health-and-social-services-reform [accessed 2021-12-09]
31. Wellbeing in South Karelia. Etelä-Karjalan sosiaali- ja terveyspiiri (EKSOTE). 2022. URL: https://eksote.fi/en/customers/ [accessed 2021-07-07]
32. Ei yhtä vaan yhteisiä ratkaisuja asiakas- ja potilastietojärjestelmien hankinnassa. Ministry of Social Affairs and Health. 2021. URL: https://soteuudistus.fi/-/1271139/ei-yhta-vaan-yhteisia-ratkaisuja-asiakas-ja-potilastietojarjestelmien-hankinnassa [accessed 2021-07-07]
33. Jormanainen V, Parhiala K, Reponen J. Highly concentrated markets of electronic health records data systems in public health centres and specialist care hospitals in 2017 in Finland. Finnish J EHealth EWelfare 2019 Mar 10;11(1-2):109-124. [doi: 10.23996/fjhw.75554]
34. Hyvä tieteellinen käytäntö ja sen loukkausepäilyjen käsitteleminen Suomessa. Tutkimuseettinen neuvottelukunta. 2021. URL: https://tenk.fi/fi/ohjeet-ja-aineistot/HTK-ohje-2012 [accessed 2021-12-13]
35. Zelmer J, Ronchi E, Hyppönen H, Lupiáñez-Villanueva F, Codagnone C, Nøhr C, et al. International health IT benchmarking: learning from cross-country comparisons. J Am Med Inform Assoc 2017 Mar 01;24(2):371-379 [FREE Full text] [doi: 10.1093/jamia/ocw111] [Medline: 27554825]
36. Bruthans J. The state of national electronic prescription systems in the EU in 2018 with special consideration given to interoperability issues. Int J Med Inform 2020 Sep;141:104205. [doi: 10.1016/j.ijmedinf.2020.104205] [Medline: 32492586]
37. eHealth Action Plan 2012-2020. European Commission. 2012. URL: https://ec.europa.eu/health/publications/ehealth-action-plan-2012-2020_en [accessed 2021-06-23]
38. Global eHealth survey 2015. World Health Organization. 2017 Jun 2. URL: https://gateway.euro.who.int/en/datasets/ehealth-survey-2015/ [accessed 2022-04-11]
39. Payne TH, Lovis C, Gutteridge C, Pagliari C, Natarajan S, Yong C, et al. Status of health information exchange: a comparison of six countries. J Glob Health 2019 Dec;9(2):0204279 [FREE Full text] [doi: 10.7189/jogh.09.020427] [Medline: 31673351]
40. Information to support well-being and service renewal. eHealth and eSocial Strategy 2020. Ministry of Social Affairs and Health (Sosiaali- ja terveysministeriö). 2015. URL: http://urn.fi/URN:ISBN:978-952-00-3575-4 [accessed 2021-05-14]

Abbreviations

CDSS: clinical decision support system
e-booking: e–appointment booking
EC: European Commission
EPR: electronic patient record
EU: European Union
GP: general practitioner
HIE: health information exchange
ICT: information and communications technology
OECD: Organization for Economic Co-operation and Development
PACS: picture archiving and communication system
RHIE: regional health information exchange
WHO: World Health Organization

Edited by C Lovis; submitted 13.12.21; peer-reviewed by B Kijsanayotin, T Heponiemi, H Yu, J Bruthans; comments to author 13.02.22; revised version received 02.06.22; accepted 03.07.22; published 12.08.22.

Please cite as:
Haverinen J, Keränen N, Tuovinen T, Ruotanen R, Reponen J
National Development and Regional Differences in eHealth Maturity in Finnish Public Health Care: Survey Study
JMIR Med Inform 2022;10(8):e35612
URL: https://medinform.jmir.org/2022/8/e35612
doi:10.2196/35612
PMID:35969462

Original Paper

Standard Vocabularies to Improve Machine Learning Model Transferability With Electronic Health Record Data: Retrospective Cohort Study Using Health Care–Associated Infection

Amber C Kiser[1], BS; Karen Eilbeck[1], MSc, PhD; Jeffrey P Ferraro[2], PhD; David E Skarda[3,4], MD; Matthew H Samore[2,5], MD; Brian Bucher[1,4], MS, MD

[1]Department of Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, UT, United States

[2]Department of Medicine, School of Medicine, University of Utah, Salt Lake City, UT, United States

[3]Center for Value-Based Surgery, Intermountain Healthcare, Salt Lake City, UT, United States

[4]Department of Surgery, School of Medicine, University of Utah, Salt Lake City, UT, United States

[5]Informatics, Decision-Enhancement and Analytic Sciences Center 2.0, Veterans Affairs Salt Lake City Health Care System, Salt Lake City, UT, United States

Corresponding Author:
Amber C Kiser, BS
Department of Biomedical Informatics
School of Medicine
University of Utah
421 Wakara Way
Suite 140
Salt Lake City, UT, 84108
United States
Phone: 1 801 581 4080
Email: amber.kiser@utah.edu

Abstract

Background: With the widespread adoption of electronic healthcare records (EHRs) by US hospitals, there is an opportunity to leverage this data for the development of predictive algorithms to improve clinical care. A key barrier in model development and implementation includes the external validation of model discrimination, which is rare and often results in worse performance. One reason why machine learning models are not externally generalizable is data heterogeneity. A potential solution to address the substantial data heterogeneity between health care systems is to use standard vocabularies to map EHR data elements. The advantage of these vocabularies is a hierarchical relationship between elements, which allows the aggregation of specific clinical features to more general grouped concepts.

Objective: This study aimed to evaluate grouping EHR data using standard vocabularies to improve the transferability of machine learning models for the detection of postoperative health care–associated infections across institutions with different EHR systems.

Methods: Patients who underwent surgery from the University of Utah Health and Intermountain Healthcare from July 2014 to August 2017 with complete follow-up data were included. The primary outcome was a health care–associated infection within 30 days of the procedure. EHR data from 0-30 days after the operation were mapped to standard vocabularies and grouped using the hierarchical relationships of the vocabularies. Model performance was measured using the area under the receiver operating characteristic curve (AUC) and F_1-score in internal and external validations. To evaluate model transferability, a difference-in-difference metric was defined as the difference in performance drop between internal and external validations for the baseline and grouped models.

Results: A total of 5775 patients from the University of Utah and 15,434 patients from Intermountain Healthcare were included. The prevalence of selected outcomes was from 4.9% (761/15,434) to 5% (291/5775) for surgical site infections, from 0.8% (44/5775) to 1.1% (171/15,434) for pneumonia, from 2.6% (400/15,434) to 3% (175/5775) for sepsis, and from 0.8% (125/15,434) to 0.9% (50/5775) for urinary tract infections. In all outcomes, the grouping of data using standard vocabularies resulted in a

reduced drop in AUC and F_1-score in external validation compared to baseline features (all P<.001, except urinary tract infection AUC: P=.002). The difference-in-difference metrics ranged from 0.005 to 0.248 for AUC and from 0.075 to 0.216 for F_1-score.

Conclusions: We demonstrated that grouping machine learning model features based on standard vocabularies improved model transferability between data sets across 2 institutions. Improving model transferability using standard vocabularies has the potential to improve the generalization of clinical prediction models across the health care system.

(JMIR Med Inform 2022;10(8):e39057) doi:10.2196/39057

KEYWORDS

standard vocabularies; machine learning; electronic health records; model transferability; data heterogeneity; machine learning

Introduction

The widespread adoption of electronic healthcare records (EHRs) by US hospitals has created an opportunity to leverage this data for the development of predictive algorithms to improve clinical care [1]. Various machine learning (ML) models have been developed to predict a variety of outcomes, including pneumonia, sepsis, and surgical site infection [2-5]. However, relatively few of these models have been implemented into clinical practice [6]. A key barrier in model development includes the validation of model discrimination across data sets [7]. Typically, validation occurs using a blind subset of data from the training data set, termed internal validation. External validation using data from a different institution is rare and often results in worse performance [8,9].

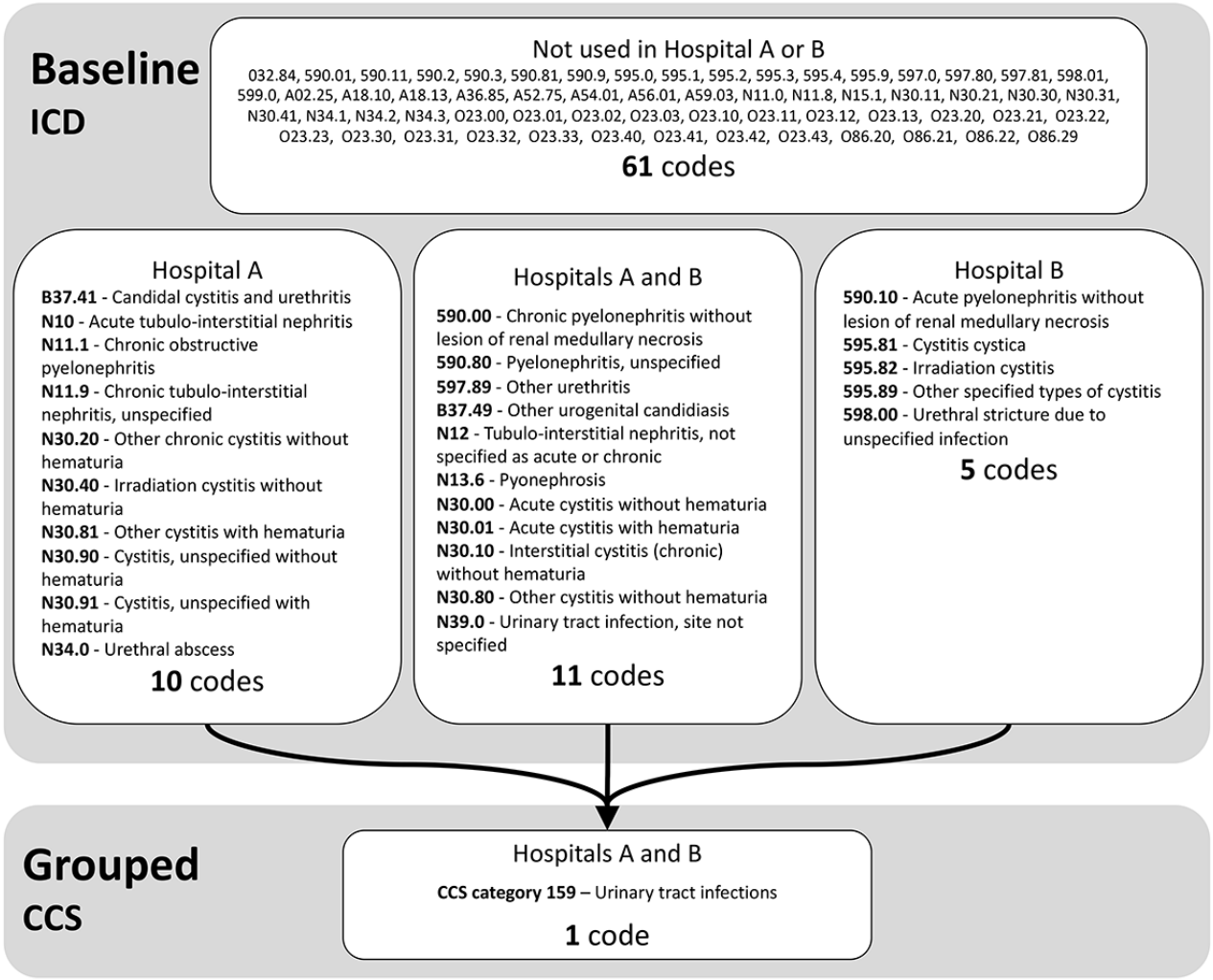
There are many reasons why ML models are not externally generalizable, including inadequate training data, overfitting of the model, and data heterogeneity [10]. With 684 different EHR vendors in the United States, the lack of interoperability between institutions, even among those with the same EHR system, substantially inhibits ML model generalizability [11]. Various methods have been proposed to improve the generalizability of ML models, including transfer learning, deep learning, and common data models (CDMs) [9,12-16]. However, data heterogeneity is an underappreciated key determinant of model transferability [17]. Data heterogeneity deriving from variation in laboratory practices, hospital medication formularies, and administrative coding practices between health care systems can impact model performance during external validation, resulting in a decreased transferability of models across institutions [18].

A solution to address the substantial data heterogeneity between health care systems is to use standard vocabularies to map EHR data elements. These vocabularies, such as the Clinical Classification Software (CCS) for International Classification of Diseases (ICD) Diagnosis Codes, Logical Observation Identifiers Names and Codes (LOINC) for health care observations, and Medi-Span for medications, can be used to support data harmonization between data sets [19-23]. The advantage of these vocabularies is a hierarchical relationship between elements, which allows the aggregation of specific clinical features to more general grouped concepts. For example, Figure 1 demonstrates how multiple ICD diagnosis codes describing “urinary tract infections” can be aggregated to 1 single CCS code. Due to variation in coding practices among health care facilities, the aggregation of concepts may improve ML model transferability during external validation.

This study’s objective was to evaluate whether aggregating EHR data elements using standard vocabularies would improve ML model transferability to an external data set. Although other works have used this method of grouping EHR data elements when developing ML models, none to our knowledge have assessed the impact of grouping on model transferability to an external data set [17]. To evaluate this objective, we classified postoperative health care–associated infections (HAIs) using EHR data from 2 independent health care systems.

HAIs pose a substantial patient safety concern, raise costs, and increase the risk of death after surgical procedures. HAIs occur in 3% to 27% of surgical patients [24,25]. Developing even 1 major postoperative complication increases a patient’s risk of postoperative mortality and readmission [26,27]. To address the challenges of HAIs, hospitals rely on surveillance programs to monitor HAI rates and develop targeted interventions to address postoperative HAIs. Hospitals that participate in quality surveillance programs reduce HAIs over time [28]. However, hospital surveillance programs rely on a manual chart review process, which is a critical barrier to the widespread adoption of surgical quality assurance programs. To overcome these difficulties, automated surveillance programs are needed to decrease the burden of the manual chart review process [29,30]. We hypothesized that ML models for HAI detection using grouped features from EHR data would improve model performance during external validation compared to ML models developed using baseline features.

Figure 1. Example of the aggregation of baseline features to grouped concepts. Multiple ICD diagnosis codes describing "urinary tract infections," including 10 used only in Hospital A, 5 used only in Hospital B, 11 used at both Hospital A and B, and 61 not used in either hospital, can be aggregated to 1 single CCS code. CCS: Clinical Classification Software; ICD: International Classification of Diseases.



Methods

Setting

We performed a retrospective cohort study using data from 2 independent health care systems: the University of Utah Health (Hospital A) with an Epic EHR and Intermountain Healthcare (Hospital B) with a Cerner EHR.

Ethics Approval

The institutional review boards at each health care system approved the study (University of Utah Health: 87482; Intermountain Healthcare: 1050851), granting a waiver of informed consent.

Data Sources, Participants, and Outcomes

Data for the study were obtained from the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) at each institution. The ACS NSQIP program is the largest surgical quality assessment program in the United States, found in over 450 hospitals [31]. As part of the program, the surgical clinical reviewers, typically nurses, are trained in NSQIP methodology and definitions [32]. NSQIP surgical clinical reviewers manually review the EHR records for all selected operative episodes to identify perioperative complications, including HAI, occurring within 30 days of the operation. All identified complications are rereviewed by the ACS surgeon champion at the participating hospital to ensure that the complications meet the ACS NSQIP definitions. Disagreements are settled when a consensus is reached, with the ACS surgeon acting as adjudicator. The interrater reliability and data quality of the NSQIP program have been previously documented [32].

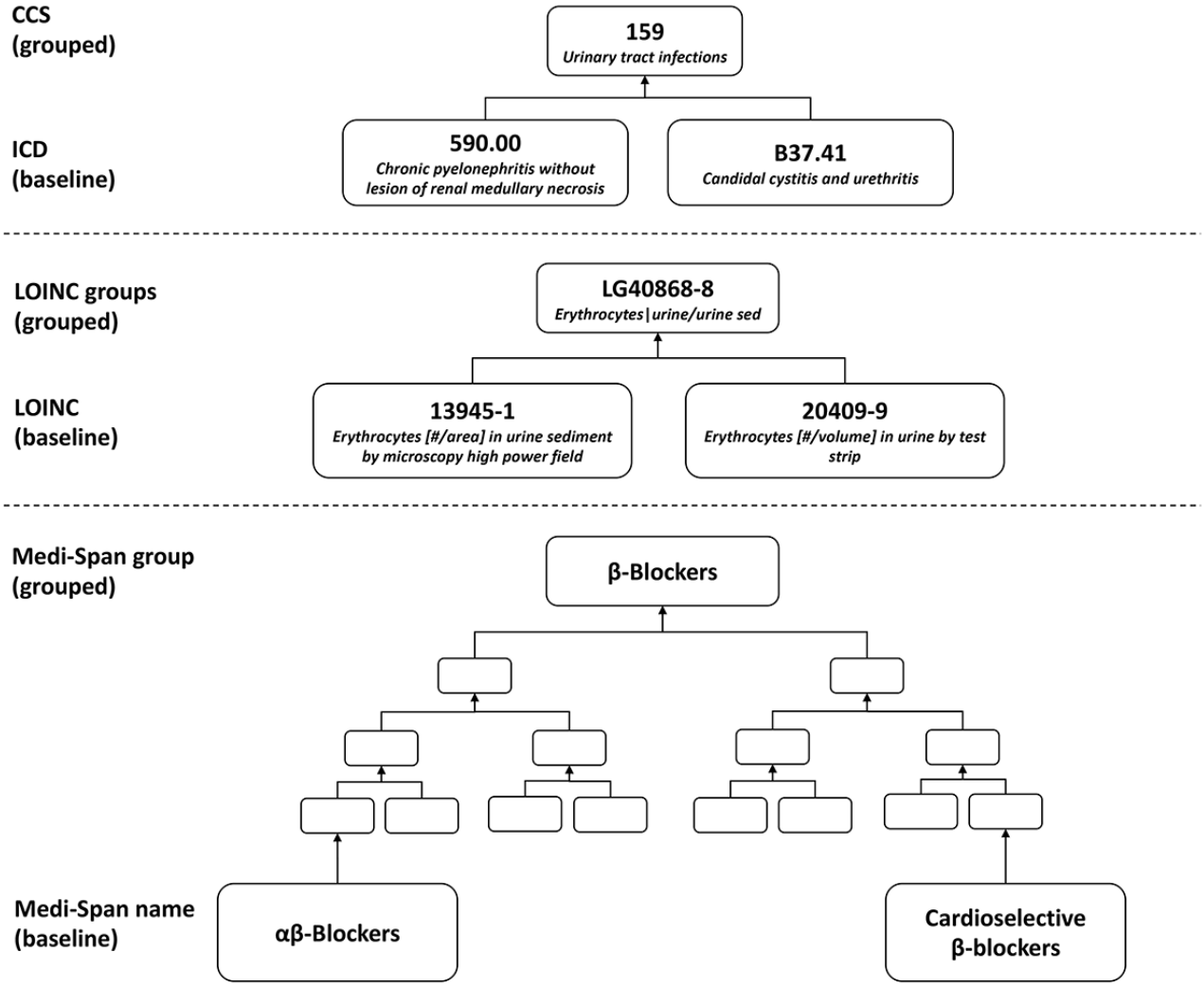
For this study, patient operative episodes were included if they underwent manual chart review as part of the ACS NSQIP program at each institution. Operative events were excluded if they had incomplete follow-up data.

The following HAIs were chosen as outcomes due to their prevalence and clinical relevance: surgical site infection (SSI), pneumonia, sepsis, and urinary tract infection (UTI). These outcomes were selected as they are the most common complications occurring after general and thoracic surgical procedures [33]. In addition, these complications are the most common underlying cause for hospital readmission after surgical procedures [27]. Cases were defined according to standard NSQIP definitions and labeled as binary values for classification.

EHR Data Element Mapping

For selected operative events, we obtained all laboratory test results, medication administration, and ICD 9th and 10th editions diagnosis codes from the EHR between 0-30 days after surgery. Although diagnosis codes are an important indicator of HAI, they often suffer from low sensitivity [34,35]. We chose to include additional clinical features, including laboratory tests and medications, to increase the sensitivity of our models. Each data category was mapped to a standard vocabulary and grouped based on the hierarchical relationships within the standard vocabularies. The Agency for Healthcare Research and Quality provides a mapping from both ICD-9 and ICD-10 codes to CCS codes in the form of a CSV file [19,20]. Diagnosis codes, represented as ICD codes in the EHR, were manually aggregated into single-level CCS codes using the CCS mapping. Laboratory test results were manually mapped to the LOINC terminology and then aggregated into LOINC groups [21,22]. Medications were automatically mapped to the Medi-Span Generic Product Identifier within the EHR [23]. In the Medi-Span hierarchy, we categorized the lowest level as baseline and the highest level as grouped. Figure 2 provides examples of aggregation for each data category. Once mapped, we created 2 discrete data sets. The baseline data set consisted of ICD codes, LOINC tests, and Medi-Span drug names. The grouped data set consisted of aggregated features, including CCS codes, LOINC groups, and Medi-Span drug groups.

Figure 2. Example of data aggregation. ICD diagnosis codes were manually aggregated into single-level CCS codes. LOINC observations were aggregated into LOINC groups, consisting of a single possible level. Medi-Span consisted of 5 different possible levels of aggregation. Medi-Span drug names were grouped into the highest level of aggregation—Medi-Span drug groups. CCS: Clinical Classification Software; ICD: International Classification of Disease; LOINC: Logical Observation Identifiers Names and Codes.



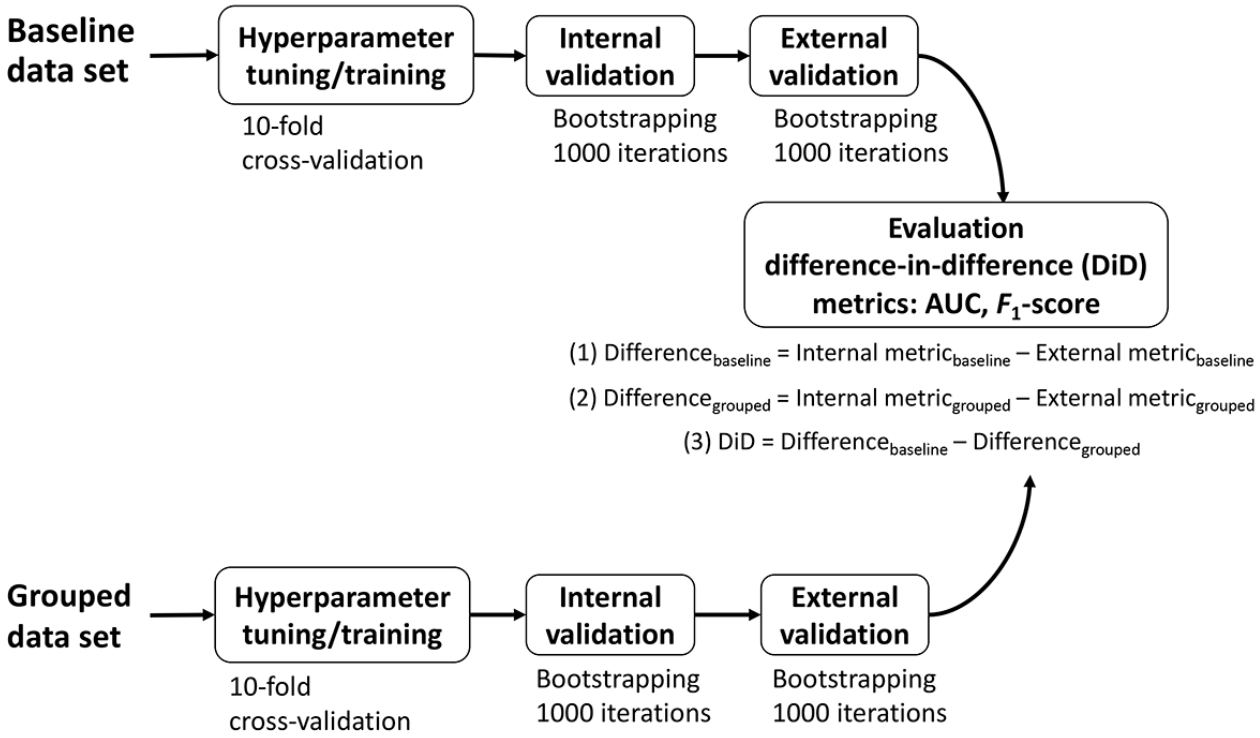
Model Development

To avoid data leakage and overfitting, we divided the data from Hospital A into hyperparameter tuning/training (70%) and internal validation (30%) data sets before preprocessing or model development. For external validation, we used 100% of the data from Hospital B. Missing data were addressed by imputing 0 for nominal variables and the median value—calculated from the training data—for continuous variables [36]. Data were standardized to have a mean of 0 and SD of 1. Figure 3 briefly describes the flow of the data through model development, validation, and final evaluation.

Separate models were developed for each outcome and data set (baseline or grouped). Each model classified whether an operative event resulted in the relevant HAI within 30 days.

Important features were identified based on the ANOVA *F*-score. Data sets with different numbers of *n*-top important features were created. In all, 4 ML algorithms were evaluated: random forest, support vector machine, logistic regression, and XGBoost [37-41]. The number of features and algorithm were included as parameters in model tuning. For each model, tuning was performed using 10-fold cross-validation to improve the internal training. The best model was selected using the area under the receiver operating characteristic curve (AUC) and F_1-score [42,43]. The final training of the models was completed using the whole training data set. To address the class imbalance, random undersampling was used during tuning within each fold of cross-validation and during final training [44]. We did not perform any balancing during validation as we wanted to test in an environment similar to real-life data where we would expect an imbalance. Model development was completed using Python software (version 3.7; Python Software Foundation) and the *scikit-learn* (version 0.22.1), *imblearn* (version 0.6.2), and *xgboost* (version 1.2.1) packages [41,45,46].

Figure 3. Flow of data through the study with the derivation for the final difference-in-difference (DiD) metric. Final evaluation steps to calculate the DiD included (1) performance difference between the internal and external validations for the baseline model; (2) performance difference between the internal and external validations for the grouped model; and (3) difference in the performance differences between the baseline and grouped models. AUC: area under the receiver operating characteristic curve.



Validation

For each model, we performed internal and external validations. For each outcome, we calculated the difference-in-difference (DiD) defined in Figure 3. DiD is a metric previously used in economics to evaluate the difference in means between 2 groups, generally a control group and an intervention group [47]. We applied it in our study to assess the difference in performance between the baseline and grouped models. A positive DiD indicates that the model developed using grouped features resulted in a reduced drop in performance during external validation compared to the model developed using baseline features.

Sensitivity Analyses

Analysis of Nonshared Codes

A separate granular data set, including baseline features but restricted to those shared by both hospital systems, was created to investigate the magnitude of performance drop in external validation attributable to nonshared codes. Training and validation were conducted as previously described. We calculated the DiD as described in Figure 3.

Analysis of Grouping Individual Categories of Data

We investigated the effect of grouping individual data categories, using only SSI, as this outcome was the most prevalent in the data. Training and validation were conducted as previously described. We compared the baseline model with models developed using data sets created with different combinations of baseline and grouped data. The combination data sets were (1) baseline diagnosis codes and laboratory tests with grouped medications, (2) baseline diagnosis codes and medications with grouped laboratory tests, and (3) grouped diagnosis codes with baseline laboratory tests and medications. We calculated the DiD as described in Figure 3.

Statistical Analysis

We performed a chi-square test of independence to determine any differences in the prevalence of the outcomes and categorical demographic variables between the institutions. For continuous demographic variables, we performed a 2-tailed,

2-sample *t* test to determine any differences between institutions. To measure model performance, bootstrapping for 1000 iterations was used to measure the mean with 95% CIs [48,49]. A 1-tailed, 1-sample *t* test was used to evaluate whether DiD metrics were significantly greater than 0. All statistical tests were completed using the *SciPy* package in Python [50].

Results

Cohort and Feature Description

A total of 5775 operative events were retrieved from Hospital A, whereas a total of 15,434 operative events were retrieved from Hospital B. Table 1 describes the study demographics.

Table 1. Study demographics for both internal and external data sets.

Characteristic	Hospital A (internal; N=5775)	Hospital B (external; N=15,434)	*P* value
Age at time of surgery (years), mean (SD)	52.6 (16.6)	53.4 (18.1)	.01
Gender, male, n (%)	2765 (47.9)	7576 (49.1)	.12
Race, n (%)			
American Indian or Alaska Native	86 (1.5)	59 (0.4)	<.001
Asian	81 (1.4)	192 (1.2)	.40
Black or African American	65 (1.1)	127 (0.8)	.05
Native Hawaiian or Pacific Islander	34 (0.6)	147 (1)	.05
White	5275 (91.3)	14,216 (92.1)	.07
Unknown or not reported	234 (4.1)	693 (4.5)	.18
Ethnicity, Hispanic, n (%)	575 (10)	1384 (9)	.03
Procedure Current Procedural Terminology code, n (%)			
0-29999 (skin/soft tissue)	968 (16.8)	2020 (13.1)	<.001
30000-39999 (cardiovascular)	594 (10.3)	2222 (14.4)	<.001
40000-49999 (gastrointestinal)	4172 (72.2)	10,796 (69.9)	.001
50000-59999 (genitourinary)	27 (0.5)	99 (0.6)	.17
60000-69999 (nervous system)	14 (0.2)	297 (1.9)	<.001
Inpatient or outpatient status, inpatient, n (%)	2831 (49)	7837 (50.8)	.02
Comorbidities, n (%)			
Diabetes mellitus	822 (14.2)	2144 (13.9)	.54
Current smoker within 1 year	799 (13.8)	2248 (14.6)	.18
Dyspnea	498 (8.6)	373 (2.4)	<.001
Functional heath status	71 (1.2)	376 (2.4)	<.001
Being ventilator-dependent	20 (0.3)	149 (1)	<.001
History of severe chronic obstructive pulmonary disease	128 (2.2)	417 (2.7)	.05
Ascites within 30 days prior to surgery	8 (0.1)	114 (0.7)	<.001
Congestive heart failure within 30 days prior to surgery	24 (0.4)	123 (0.8)	.004
Hypertension requiring medication	1940 (33.6)	5455 (35.3)	.02
Acute renal failure	9 (0.2)	53 (0.3)	.03
Currently requiring or on dialysis	100 (1.7)	283 (1.8)	.66
Disseminated cancer	187 (3.2)	246 (1.6)	<.001
Open wound with or without infection	287 (5)	512 (3.3)	<.001
Steroid or immunosuppressant use for chronic condition	351 (6.1)	644 (4.2)	<.001
>10% loss of body weight in the 6 months prior to surgery	145 (2.5)	372 (2.4)	.71
Bleeding disorder	151 (2.6)	1013 (6.6)	<.001

Table 2 describes the prevalence of HAI outcomes within each institution. There were no significant differences in the

prevalence of SSI (*P*=.77), sepsis (*P*=.09), or UTI (*P*=.75). The prevalence of pneumonia was significantly higher (*P*=.03) in Hospital B.

Table 2. Prevalence of selected outcomes in each hospital system.

Outcome	Hospital A (N=5775), n (%)	Hospital B (N=15,434), n (%)	*P* value
Surgical site infection	291 (5)	761 (4.9)	.77
Pneumonia	44 (0.8)	171 (1.1)	.03[a]
Sepsis	175 (3)	400 (2.6)	.09
Urinary tract infection	50 (0.9)	125 (0.8)	.75

[a]Pneumonia was significantly more prevalent in Hospital B (*P*<.05).

Model Development and Validation

DiD metrics are reported in Table 3. Tables S1 and S2 in Multimedia Appendix 1 detail the selected model parameters. Model calibration can be found in Table S3 and Figures S1-S4 in Multimedia Appendix 1. Standards for Reporting Diagnostic Accuracy Studies flow diagrams of patient data through the top-performing models can be seen in Figures S5-S16 in Multimedia Appendix 1.

After external validation, all models produced significantly positive AUC and F_1-score DiDs when comparing the performance of the baseline and grouped models (all *P*<.001, except UTI AUC: *P*=.002). A forest plot in Figure S17 in Multimedia Appendix 1 illustrates the AUC and F_1-score DiDs.

Table 3. Difference-in-difference (DiD) metrics for each outcome. Means are based on 1000 bootstrapped iterations with 95% CIs. A positive DiD indicates that the grouped model resulted in a reduced drop in performance compared to the baseline model.

Outcome, metric	Top baseline algorithm	Top grouped algorithm	Baseline internal validation, mean (95% CI)	Baseline external validation, mean (95% CI)	Grouped internal validation, mean (95% CI)	Grouped external validation, mean (95% CI)	DiD, mean (95% CI)	*P* value
SSI[a]	SVM[b]	LR[c]						
AUC[d]			0.906 (0.904-0.908)	0.763 (0.762-0.764)	0.904 (0.903-0.906)	0.833 (0.833-0.834)	0.072 (0.070-0.074)	<.001
F_1-score			0.501 (0.499-0.503)	0.300 (0.299-0.302)	0.476 (0.474-0.478)	0.376 (0.375-0.376)	0.100 (0.097-0.103)	<.001
Pneumonia	LR	SVM						
AUC			0.953 (0.949-0.957)	0.683 (0.682-0.685)	0.994 (0.994-0.995)	0.973 (0.973-0.974)	0.250 (0.247-0.252)	<.001
F_1-score			0.504 (0.498-0.509)	0.302 (0.299-0.305)	0.456 (0.452-0.461)	0.467 (0.465-0.468)	0.212 (0.206-0.218)	<.001
Sepsis	LR	RF[e]						
AUC			0.964 (0.963-0.964)	0.890 (0.889-0.891)	0.948 (0.946-0.949)	0.883 (0.883-0.884)	0.008 (0.007-0.010)	<.001
F_1-score			0.469 (0.467-0.472)	0.050 (0.050-0.050)	0.419 (0.416-0.422)	0.092 (0.092-0.093)	0.091 (0.089-0.093)	<.001
UTI[f]	SVM	LR						
AUC			0.898 (0.895-0.900)	0.886 (0.885-0.887)	0.936 (0.934-0.939)	0.929 (0.928-0.930)	0.006 (0.002-0.009)	.002
F_1-score			0.153 (0.148-0.158)	0.063 (0.061-0.064)	0.244 (0.241-0.246)	0.225 (0.224-0.226)	0.073 (0.068-0.077)	<.001

[a]SSI: surgical site infection.

[b]SVM: support vector machine.

[c]LR: logistic regression.

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

[e]RF: random forest.

[f]UTI: urinary tract infection.

Sensitivity Analyses

Effect of Nonshared Codes

Table 4 describes the EHR data elements shared between hospitals. We found that 44.8% (4284/9559) of baseline features present in the training set were not present in the external set, whereas all grouped features present in the training set were present in the external set.

After external validation, all models, except UTI (P=.002), produced significantly positive AUC DiDs (all P<.001) when comparing the performance of the baseline and granular models. All outcomes produced significantly positive F_1-score DiDs (all P<.001) when comparing the performance of the baseline and granular models.

The magnitude of the AUC and F_1-score DiDs calculated from the comparison of the baseline and grouped models were greater than those calculated from the comparison of the baseline and granular models in all outcomes, except the AUC DiD for sepsis, as represented in Table 5. Full internal and external validation results can be found in Table S4 in Multimedia Appendix 1.

Table 4. Number of features in each category (diagnosis, medication, and laboratory) for Hospital A, Hospital B, and those shared between them.

Features	Training Set (Hospital A), n	External Set (Hospital B), n	Shared, n
Baseline			
Total	9559	7926	5275
ICD[a] diagnosis codes	7708	6859	4392
Medi-Span drug names	1311	531	531
LOINC[b] codes	540	536	352
Grouped			
Total	805	817	805
CCS[c] diagnosis codes	287	287	287
Medi-Span drug groups	94	94	94
LOINC groups	424	436	424

[a]ICD: International Classification of Diseases.

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

[c]CCS: Clinical Classification Software.

Table 5. Difference-in-difference (DiD) metrics for the comparison between baseline and granular models and the comparison between baseline and grouped models. A positive DiD indicates the comparison model resulted in a reduced drop in performance compared to the baseline model.

Metric, outcome	Granular comparison, DiD (95% CI)	Grouped comparison, DiD (95% CI)
AUC[a]		
SSI[b]	0.035 (0.033-0.037)	0.072 (0.070-0.074)
Pneumonia	0.226 (0.223-0.229)	0.250 (0.247-0.252)
Sepsis	0.015 (0.013-0.017)	0.008 (0.007-0.010)
UTI[c]	–0.049 (–0.052 to –0.045)	0.006 (0.002-0.009)
F_1-score		
SSI	0.017 (0.014-0.020)	0.100 (0.097-0.103)
Pneumonia	0.186 (0.179-0.193)	0.212 (0.206-0.218)
Sepsis	0.026 (0.023-0.028)	0.091 (0.089-0.093)
UTI	0.039 (0.035-0.043)	0.073 (0.068-0.077)

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

[b]SSI: surgical site infection.

[c]UTI: urinary tract infection.

Effect of Grouping Individual Categories of Data

In the second sensitivity analysis, all AUC and F_1-score DiDs were significantly positive (all P<.001) when comparing the performance of the baseline and combination models, as displayed in Table 6. The combination model with grouped medications, Combination 1, resulted in the greatest AUC DiD.

The combination model with grouped diagnosis codes, Combination 3, resulted in the greatest F_1-score DiD.

Table 6. Comparison of models developed from baseline data with models developed from the combination of baseline and grouped data. The difference-in-difference (DiD) reflects the AUC and F_1-score for surgical site infection. A positive DiD indicates the combination model resulted in a smaller drop in performance than the baseline model.

Combination	Medications	Laboratory tests	Diagnosis codes	AUC[a], DiD (95% CI)	F_1-score, DiD (95% CI)	*P* value
Combination 1	Grouped	Baseline	Baseline	0.054 (0.052-0.057)	0.072 (0.069-0.074)	<.001
Combination 2	Baseline	Grouped	Baseline	0.012 (0.010-0.014)	0.046 (0.043-0.049)	<.001
Combination 3	Baseline	Baseline	Grouped	0.049 (0.047-0.051)	0.134 (0.131-0.137)	<.001

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

Discussion

We investigated the effect that grouping EHR data using standard vocabularies has on ML model transferability during external validation. There are several novel and significant findings of our work. First, ML models for HAI detection with grouped features based on standard vocabularies resulted in a reduced drop in performance when validated on an external data set compared to baseline features. Second, there was significant heterogeneity of EHR data elements between health care systems, as 45% of data elements present in the training set were not present in the external set. Third, ML models developed from grouped data sets resulted in greater performance gains after external validation compared to data sets restricted to shared codes alone. Lastly, we found that grouping diagnosis codes and medications was important to model transferability when compared to laboratory tests.

We demonstrated that grouping features using standard vocabularies improved model transferability during external validation. We found on average a 51% decrease and 65% decrease in the performance drop of AUC and F_1-score, respectively, during external validation when using grouped data compared to baseline data. This improvement in transferability can be attributed to better syntactic and semantic interoperability. Using grouped features allows the model to overcome the challenges of data heterogeneity, such as differences in coding practice and hospital formularies, that arise when using granular codes. A single feature from the grouped model can represent several distinct features from the baseline model (Figure 1). Hence, this method can generalize to an unknown data set as no knowledge of the future data set is required when selecting features or training the model. Although the practice of grouping features is common, our study is novel in that to our knowledge, previous studies have not evaluated model transferability in an external data set when grouping features based on standard vocabularies.

The data heterogeneity seen in our data highlights the difficulty when creating generalizable ML models. Shared codes accounted for 57% (4392/7708) of the ICD diagnosis codes used in Hospital A and 64% (4392/6859) of the ICD diagnosis codes used in Hospital B. To our knowledge, none have compared ICD code usage between hospitals. For several common conditions, there are numerous ICD diagnosis codes available. For example, diabetes mellitus type II has 56 ICD-9 and ICD-10 codes available [51,52]. Variation in coding practices between health care systems can result in several individual codes not being present in a given data set. Differences in laboratory practices or hospital medication formularies may also contribute to EHR data heterogeneity. Extensive feature engineering is typically performed to overcome this challenge before model development [53]. Feature engineering, while creating highly relevant features for the given use case, represents a substantial barrier to model generalizability. Our study demonstrated that grouping features can overcome challenges created by data heterogeneity.

In the first sensitivity analysis, we found that although models developed with granular data sets restricted to shared codes resulted in a reduced drop in performance when compared to a baseline model, models developed from grouped data sets resulted in an even smaller drop in performance. The models developed using grouped data sets resulted in an additional 41% decrease and 70% decrease in performance drop of AUC and F_1-score, respectively, during external validation on average. These results provide further evidence that grouping features using standard vocabularies produces greater benefits than just restricting features to those shared by other hospital systems.

In the second sensitivity analysis, we found that the most important factors when improving transferability included grouping both diagnosis codes and medications. This result could be explained by the amount of information lost due to variation in coding practices and prescription preferences when using baseline data. Rasmy et al [54] compared models using different representations of diagnosis codes in the EHR. The study found that models developed with data mapped to the Unified Medical Language System (UMLS) produced the highest AUC, whereas models developed with data mapped to CCS codes produced the lowest AUC. However, this previous study did not have an external data set to compare performance.

Other studies have used various methods to improve model transferability, including transfer learning, deep learning, and anchor learning [9,12-16]. Curth et al [12] found that using transfer learning significantly increased model performance, where the AUC increased as much as from 4.7% to 7.0% depending on the use case. Although transfer learning has been shown to be successful, it requires models to be trained with data from the internal and external sites. Rasmy et al [15] found an average drop of 3.6% in AUC when evaluating the generalizability of a recurrent neural network. In our study, we found the average drop in AUC to be 13% in models developed using baseline data but only 4% in models developed using

grouped data. Kashyap et al [13] found performance drops in both recall and precision when validating the model at an external site after using anchor learning. Our study evaluated a method to achieve comparable model transferability without requiring any knowledge of the external site or a deep learning model.

Mapping data to CDMs can facilitate the sharing of data and models across institutions as seen in several recent studies [13,55]. Recent work, such as that from Tian et al [9], has built frameworks for model sharing and generalizability that use CDMs in their pipeline [17]. The use of a CDM involves mapping data to standard vocabularies as we did in our study, which addresses the problem of syntax by standardizing the vocabulary. In our study, we further address the problem of semantics, where different hospitals may use the same vocabulary, but coding practices may result in different codes representing the same condition.

We acknowledge several limitations to this study. Our use case consisted of HAI detection in patients who underwent surgery. The benefit of grouping feature sets for ML development may not be consistent across other use cases. We only used EHR data elements for which there are standard vocabularies available, excluding features such as microbiology reports or clinical text. It is likely that including these additional features would improve ML model performance at the expense of requiring an extensive amount of feature engineering. We used Medi-Span, a proprietary vocabulary, as both hospital EHRs mapped medications to this system. Other vocabularies, such as RxNorm, could be used. There are several different terminologies that can be used to group diagnosis codes in addition to CCS, including UMLS, as was studied by Rasmy et al [54]. Their work indicates that using UMLS to group diagnosis codes could produce an even smaller drop in performance than we found with CCS. This method would be a valuable investigation for future studies that could lead to even greater results. The terminologies and levels chosen for our study could be modified for different use cases.

This study has substantial implications for the application of ML models to clinical practice. Significant improvements in patient care can be achieved with ML models as demonstrated in previous studies [13,14,56,57]. However, external validation remains one of the most serious barriers to the widespread use of ML models in clinical practice [6,58]. We found that 2 independent hospitals only shared 55% of baseline EHR data elements, highlighting the difficulty when creating generalizable ML models. Current practices to overcome the data heterogeneity between data sets involve extensive feature engineering, which is burdensome during model deployment at a new health care system where EHR data elements are not mapped to a CDM [59]. We demonstrated the novel finding that grouping features with standard vocabularies can overcome the challenge of data heterogeneity and improve ML model performance in external data sets. The method of grouping features based on standard vocabularies will improve the transferability of models, allowing for more widespread use of these ML models between health care systems.

Acknowledgments

This research was supported by a training grant (T15LM007124) from the National Library of Medicine (ACK) and a grant (1K08HS025776) from the Agency for Healthcare Research and Quality (BB). The computational resources used were partially funded by the National Institutes of Health Shared Instrumentation (grant 1S10OD021644-01A1). The National Institutes of Health and the Agency for Healthcare Research and Quality had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

The code for this article can be found in the public GitHub repository amberkiser/MachineLearningTransferability. The data underlying this article cannot be shared publicly as it contains protected health information.

Authors' Contributions

ACK contributed to methodology; writing the code and performing the analysis (software); formal analysis; and writing—original draft, review, and editing. KE contributed to writing—original draft, review, and editing—and supervision. JPF provided resources and contributed to data curation and writing—review and editing. DES provided resources and contributed to data curation and writing—review and editing. MHS provided resources and contributed to data curation and writing—review and editing. BB contributed to conceptualization; methodology; writing—original draft, review, and editing; and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables and figures.
[DOCX File , 1175 KB - medinform_v10i8e39057_app1.docx]

References

1. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014 Dec;52:28-35 [FREE Full text] [doi: 10.1016/j.jbi.2014.02.003] [Medline: 24534443]
2. Meystre S, Gouripeddi R, Tieder J, Simmons J, Srivastava R, Shah S. Enhancing comparative effectiveness research with automated pediatric pneumonia detection in a multi-institutional clinical repository: a PHIS+ pilot study. J Med Internet Res 2017 May 15;19(5):e162 [FREE Full text] [doi: 10.2196/jmir.6887] [Medline: 28506958]
3. Ge Y, Wang Q, Wang L, Wu H, Peng C, Wang J, et al. Predicting post-stroke pneumonia using deep neural network approaches. Int J Med Inform 2019 Dec;132:103986 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.103986] [Medline: 31629312]
4. 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]
5. Chen W, Lu Z, You L, Zhou L, Xu J, Chen K. Artificial intelligence-based multimodal risk assessment model for surgical site infection (AMRAMS): development and validation study. JMIR Med Inform 2020 Jun 15;8(6):e18186 [FREE Full text] [doi: 10.2196/18186] [Medline: 32538798]
6. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015 Jan;68(1):25-34. [doi: 10.1016/j.jclinepi.2014.09.007] [Medline: 25441703]
7. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. JAMA 2018 Jul 03;320(1):27-28. [doi: 10.1001/jama.2018.5602] [Medline: 29813156]
8. 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]
9. Tian Y, Chen W, Zhou T, Li J, Ding K, Li J. Establishment and evaluation of a multicenter collaborative prediction model construction framework supporting model generalization and continuous improvement: a pilot study. Int J Med Inform 2020 Sep;141:104173. [doi: 10.1016/j.ijmedinf.2020.104173] [Medline: 32531725]
10. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016 Jan;69:245-247 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
11. Office of the National Coordinator for Health Information Technology. Certified health IT developers and editions reported by health care professionals participating in the Medicare EHR Incentive Program. HealthIT. 2017 Jul. URL: https://www.healthit.gov/data/quickstats/health-care-professional-health-it-developers [accessed 2021-12-07]
12. Curth A, Thoral P, van den Wildenberg W, Bijlstra P, de Bruin D, Elbers P, et al. Transferring clinical prediction models across hospitals and electronic health record systems. 2020 Mar 28 Presented at: ECML PKDD 2019: Machine Learning and Knowledge Discovery in Databases; September 16-20, 2019; Würzburg, Germany p. 605-621. [doi: 10.1007/978-3-030-43823-4_48]
13. Kashyap M, Seneviratne M, Banda JM, Falconer T, Ryu B, Yoo S, et al. Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. J Am Med Inform Assoc 2020 Jun 01;27(6):877-883 [FREE Full text] [doi: 10.1093/jamia/ocaa032] [Medline: 32374408]
14. Hassanzadeh H, Nguyen A, Karimi S, Chu K. Transferability of artificial neural networks for clinical document classification across hospitals: a case study on abnormality detection from radiology reports. J Biomed Inform 2018 Sep;85:68-79 [FREE Full text] [doi: 10.1016/j.jbi.2018.07.017] [Medline: 30026067]
15. Rasmy L, Wu Y, Wang N, Geng X, Zheng WJ, Wang F, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. J Biomed Inform 2018 Aug;84:11-16 [FREE Full text] [doi: 10.1016/j.jbi.2018.06.011] [Medline: 29908902]
16. Chi S, Li X, Tian Y, Li J, Kong X, Ding K, et al. Semi-supervised learning to improve generalizability of risk prediction models. J Biomed Inform 2019 Apr;92:103117 [FREE Full text] [doi: 10.1016/j.jbi.2019.103117] [Medline: 30738948]
17. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018 Aug 01;25(8):969-975 [FREE Full text] [doi: 10.1093/jamia/ocy032] [Medline: 29718407]
18. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J 2021 Jan;14(1):49-58 [FREE Full text] [doi: 10.1093/ckj/sfaa188] [Medline: 33564405]
19. Clinical Classification Software (CCS) for ICD-9-CM. Healthcare Cost and Utilization Project. 2017 Mar. URL: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp [accessed 2020-06-23]
20. Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses. Healthcare Cost and Utilization Project. 2022 Feb. URL: https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp [accessed 2022-07-19]
21. LOINC from Regenstrief. LOINC. URL: https://loinc.org/ [accessed 2020-06-23]
22. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clin Chem 1996 Jan;42(1):81-90. [Medline: 8565239]

23. Medi-Span: power up your drug decisions with smart data. Wolters Kluwer. URL: https://www.wolterskluwer.com/en/solutions/medi-span [accessed 2020-06-23]
24. Bucher BT, Ferraro JP, Finlayson SRG, Chapman WW, Gundlapalli AV. Use of computerized provider order entry events for postoperative complication surveillance. JAMA Surg 2019 Apr 01;154(4):311-318 [FREE Full text] [doi: 10.1001/jamasurg.2018.4874] [Medline: 30586132]
25. de Vries EN, Prins HA, Crolla RMPH, den Outer AJ, van Andel G, van Helden SH, SURPASS Collaborative Group. Effect of a comprehensive surgical safety system on patient outcomes. N Engl J Med 2010 Nov 11;363(20):1928-1937. [doi: 10.1056/NEJMsa0911535] [Medline: 21067384]
26. Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. N Engl J Med 2009 Oct 01;361(14):1368-1375. [doi: 10.1056/NEJMsa0903048] [Medline: 19797283]
27. Merkow RP, Ju MH, Chung JW, Hall BL, Cohen ME, Williams MV, et al. Underlying reasons associated with hospital readmission following surgery in the United States. JAMA 2015 Feb 03;313(5):483-495. [doi: 10.1001/jama.2014.18614] [Medline: 25647204]
28. Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. Ann Surg 2009 Sep;250(3):363-376. [doi: 10.1097/SLA.0b013e3181b4148f] [Medline: 19644350]
29. Shi J, Liu S, Pruitt LCC, Luppens CL, Ferraro JP, Gundlapalli AV, et al. Using natural language processing to improve EHR structured data-based surgical site infection surveillance. AMIA Annu Symp Proc 2019;2019:794-803 [FREE Full text] [Medline: 32308875]
30. Zhu Y, Simon GJ, Wick EC, Abe-Jones Y, Najafi N, Sheka A, et al. Applying machine learning across sites: external validation of a surgical site infection detection algorithm. J Am Coll Surg 2021 Jun;232(6):963-971.e1 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2021.03.026] [Medline: 33831539]
31. Ko CY, Hall BL, Hart AJ, Cohen ME, Hoyt DB. The American College of Surgeons National Surgical Quality Improvement Program: achieving better and safer surgery. Jt Comm J Qual Patient Saf 2015 May;41(5):199-204. [doi: 10.1016/s1553-7250(15)41026-8] [Medline: 25977246]
32. Shiloach M, Frencher SK, Steeger JE, Rowell KS, Bartzokis K, Tomeh MG, et al. Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. J Am Coll Surg 2010 Jan;210(1):6-16. [doi: 10.1016/j.jamcollsurg.2009.09.031] [Medline: 20123325]
33. Dencker EE, Bonde A, Troelsen A, Varadarajan KM, Sillesen M. Postoperative complications: an observational study of trends in the United States from 2012 to 2018. BMC Surg 2021 Nov 06;21(1):393 [FREE Full text] [doi: 10.1186/s12893-021-01392-z] [Medline: 34740362]
34. van Mourik MSM, van Duijn PJ, Moons KGM, Bonten MJM, Lee GM. Accuracy of administrative data for surveillance of healthcare-associated infections: a systematic review. BMJ Open 2015 Aug 27;5(8):e008424 [FREE Full text] [doi: 10.1136/bmjopen-2015-008424] [Medline: 26316651]
35. Redondo-González O, Tenías JM, Arias Á, Lucendo AJ. Validity and reliability of administrative coded data for the identification of hospital-acquired infections: an updated systematic review with meta-analysis and meta-regression analysis. Health Serv Res 2018 Jun;53(3):1919-1956 [FREE Full text] [doi: 10.1111/1475-6773.12691] [Medline: 28397261]
36. Hu Z, Melton GB, Arsoniadis EG, Wang Y, Kwaan MR, Simon GJ. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. J Biomed Inform 2017 Apr;68:112-120 [FREE Full text] [doi: 10.1016/j.jbi.2017.03.009] [Medline: 28323112]
37. Breiman L. Random forests. Mach Learn 2001 Oct;45:5-32. [doi: 10.1023/A:1010933404324]
38. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011 Apr;2(3):1-27. [doi: 10.1145/1961189.1961199]
39. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 1999 Mar 26;10(3):61-74 [FREE Full text]
40. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med 2011 Oct;18(10):1099-1104 [FREE Full text] [doi: 10.1111/j.1553-2712.2011.01185.x] [Medline: 21996075]
41. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Aug 13 Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
42. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
43. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. 2005 Presented at: ECIR 2005: Advances in Information Retrieval; March 21-23, 2005; Santiago de Compostela, Spain p. 345-359. [doi: 10.1007/978-3-540-31865-1_25]
44. Hasanin T, Khoshgoftaar T. The effects of random undersampling with simulated class imbalance for big data. 2018 Aug 06 Presented at: 2018 IEEE International Conference on Information Reuse and Integration (IRI); 06-09 July, 2018; Salt Lake City, UT p. 6-9. [doi: 10.1109/iri.2018.00018]

45. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017 Jan;18(1):559-563 [FREE Full text]
46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011 Oct;12:2825-2830 [FREE Full text]
47. Schwerdt G, Woessmann L. Chapter 1 - Empirical methods in the economics of education. In: Bradley S, Green C, editors. The Economics of Education. 2nd ed. Cambridge, MA: Academic Press; 2020:3-20.
48. Margolis D, Bilker W, Boston R, Localio R, Berlin J. Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. J Clin Epidemiol 2002 May;55(5):518-524. [doi: 10.1016/s0895-4356(01)00512-1] [Medline: 12007556]
49. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J Anal Test 2018;2(3):249-262 [FREE Full text] [doi: 10.1007/s41664-018-0068-2] [Medline: 30842888]
50. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar 3;17(3):261-272 [FREE Full text] [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
51. 2014 ICD-9-CM Diagnosis Codes: diabetes mellitus 250. ICD9Data. URL: http://www.icd9data.com/2014/Volume1/240-279/249-259/250/default.htm [accessed 2020-06-22]
52. ICD-10-CM Section E08-E13: diabetes mellitus. ICD.Codes. URL: https://icd.codes/icd10cm/chapter4/E08-E13 [accessed 2020-06-22]
53. Romero-Brufau S, Whitford D, Johnson MG, Hickman J, Morlan BW, Therneau T, et al. Using machine learning to improve the accuracy of patient deterioration predictions: Mayo Clinic Early Warning Score (MC-EWS). J Am Med Inform Assoc 2021 Jun 12;28(6):1207-1215 [FREE Full text] [doi: 10.1093/jamia/ocaa347] [Medline: 33638343]
54. Rasmy L, Tiryaki F, Zhou Y, Xiang Y, Tao C, Xu H, et al. Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. J Am Med Inform Assoc 2020 Oct 01;27(10):1593-1599 [FREE Full text] [doi: 10.1093/jamia/ocaa180] [Medline: 32930711]
55. Jin S, Kostka K, Posada JD, Kim Y, Seo SI, Lee DY, et al. Prediction of major depressive disorder following beta-blocker therapy in patients with cardiovascular diseases. J Pers Med 2020 Dec 18;10(4):288 [FREE Full text] [doi: 10.3390/jpm10040288] [Medline: 33352870]
56. Le S, Hoffman J, Barton C, Fitzgerald JC, Allen A, Pellegrini E, et al. Pediatric severe sepsis prediction using machine learning. Front Pediatr 2019;7:413 [FREE Full text] [doi: 10.3389/fped.2019.00413] [Medline: 31681711]
57. Lindberg DS, Prosperi M, Bjarnadottir RI, Thomas J, Crane M, Chen Z, et al. Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using EHR and electronic administrative data: a machine-learning approach. Int J Med Inform 2020 Nov;143:104272 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104272] [Medline: 32980667]
58. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
59. Garza M, del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]

Abbreviations

ACS: American College of Surgeons
AUC: area under the receiver operating characteristic curve
CCS: Clinical Classification Software
CDM: common data model
DiD: difference-in-difference
EHR: electronic healthcare records
HAI: health care–associated infection
ICD: International Classification of Diseases
LOINC: Logical Observation Identifiers Names and Codes
ML: machine learning
NSQIP: National Surgical Quality Improvement Program
SSI: surgical site infection
UMLS: Unified Medical Language System4^
UTI: urinary tract infection

Edited by C Lovis; submitted 27.04.22; peer-reviewed by Z Ren, Y Xu; comments to author 15.07.22; revised version received 09.08.22; accepted 15.08.22; published 30.08.22.

Please cite as:
Kiser AC, Eilbeck K, Ferraro JP, Skarda DE, Samore MH, Bucher B
Standard Vocabularies to Improve Machine Learning Model Transferability With Electronic Health Record Data: Retrospective Cohort Study Using Health Care–Associated Infection
JMIR Med Inform 2022;10(8):e39057
URL: https://medinform.jmir.org/2022/8/e39057
doi:10.2196/39057
PMID:36040784

Original Paper

Implementing Electronic Health Records in Primary Care Using the Theory of Change: Nigerian Case Study

Taiwo Adedeji[1*], BSc, MSc, MRES, PhD; Hamish Fraser[2*], MBChB, MSc; Philip Scott[3*], MSc, PhD

[1]School of Computing, University of Portsmouth, Portsmouth, United Kingdom

[2]Brown Center for Biomedical Informatics, Brown University, Providence, RI, United States

[3]Institute of Management and Health, University of Wales Trinity Saint David, Carmarthen, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Taiwo Adedeji, BSc, MSc, MRES, PhD
School of Computing, University of Portsmouth
Buckingham Building, Lion Terrace
Portsmouth, PO1 3HE
United Kingdom
Phone: 44 2392846429
Email: taiwo.adedeji@port.ac.uk

Abstract

Background: Digital health has been a tool of transformation for the delivery of health care services globally. An electronic health record (EHR) system can solve the bottleneck of paper documentation in health service delivery if it is successfully implemented, but poor implementation can lead to a waste of resources. The study of EHR system implementation in low- and middle-income countries (LMICs) is of particular interest to health stakeholders such as policy makers, funders, and care providers because of the efficiencies and evidence base that could result from the appropriate evaluation of such systems.

Objective: We aimed to develop a theory of change (ToC) for the implementation of EHRs for maternal and child health care delivery in LMICs. The ToC is an outcomes-based approach that starts with the long-term goals and works backward to the inputs and mediating components required to achieve these goals for complex programs.

Methods: We used the ToC approach for the whole implementation's life cycle to guide the pilot study and identify the preconditions needed to realize the study's long-term goal at Festac Primary Health Centre in Lagos, Nigeria. To evaluate the maturity of the implementation, we adapted previously defined success factors to supplement the ToC approach.

Results: The initial ToC map showed that the long-term goal was an improved service delivery in primary care with the introduction of EHRs. The revised ToC revealed that the long-term change was the improved maternal and child health care delivery at Festac Primary Health Center using EHRs. We proposed a generic ToC map that implementers in LMICs can use to introduce an optimized EHR system, with assumptions about sustainability and other relevant factors. The outcomes from the critical success factors were sustainability: the sustained improvements included trained health care professionals, a change in mindset from using paper systems toward digital health transformation, and using the project's laptops to collect aggregate data for the District Health Information System 2–based national health information management system; financial: we secured funding to procure IT equipment, including servers, laptops, and networking, but the initial cost of implementation was high, and funds mainly came from the funding partner; and organizational: the health professionals, especially the head of nursing and health information officers, showed significant commitment to adopting the EHR system, but certain physicians and midwives were unwilling to use the EHR system initially until they were persuaded or incentivized by the management.

Conclusions: This study shows that the ToC is a rewarding approach to framing dialogue with stakeholders and serves as a framework for planning, evaluation, learning, and reflection. We hypothesized that any future health IT implementation in primary care could adapt our ToC approach to their contexts with necessary modifications based on inherent characteristics.

(JMIR Med Inform 2022;10(8):e33491) doi:10.2196/33491

KEYWORDS

theory of change; electronic health records; maternal and child health; primary health center; success criteria

Introduction

Background

Globally, digital health has been a tool of transformation for the delivery of health care services [1]. There is a plethora of health records in paper format resulting from the handling of clinical documentation across health care facilities in low- and middle-income countries (LMICs). In LMICs, few electronic health records (EHRs) exist at public primary health centers (PHCs), the first entry point for citizens or patients seeking essential health care services [2-4]. An EHR is defined as "a repository of information regarding the health status of a subject of care, in computer processable form" [5]. An EHR system can solve the bottleneck of paper documentation in health service delivery if it is successfully implemented, but poor implementation can lead to a waste of resources [6]. The study of EHR system implementation in LMICs is of particular interest to health stakeholders such as policy makers, funding agencies, and care providers because of the efficiencies and evidence base that could result from the appropriate evaluation of such systems [7]. Some progress has already been made regarding EHR implementation in LMICs, but sustainability and widespread adoption remain elusive [8,9]. A few examples of such developments in health care improvements include efficiency gains (such as quicker and more accurate reporting, reduced duplication of documentation, and quicker access to patients' records), better patient tracking (such as immunization records and clinic attendance), and mobile health apps (ubiquitous access to remote care for patients) [10], an example of which is Virtual Doctors, a UK-based charity that specializes in telemedicine and provides remote medical advice to local health workers to reduce unnecessary hospital referrals. Currently, the charity is working with PHCs in Zambia and Malawi, where volunteer physicians, mostly from the United Kingdom, provide medical support through a mobile app. These volunteer physicians provide medically qualified advice where the local community only has a community health worker, leading to faster diagnosis and treatment [11]. Another example is iSanté, Haiti's national electronic medical record system. This EHR system was implemented in 100 sites across Haiti primarily to support the delivery of the national HIV program [12]; it also supports antenatal care (ANC), delivery, and essential primary care services.

Maternal and Child Health Care in Nigerian Primary Health Care

A significant health need in Nigerian primary care, in common with primary care in other LMICs, is maternal and child health care (MCH) [7]. Women and their children would usually attend the health facility for ANC, delivery, immunization, and family planning services [13]. It is essential to manage the health records of these patients or citizens effectively and efficiently to ensure effective clinical workflow and patient safety. Although paper-based health records seem to be structured in supporting care delivery, EHRs prove to be more consistent, readily available, and scalable for continuity of care [14].

EHR Implementation in Nigeria

In high-income countries (HICs), there has been widespread adoption of EHRs, but this is not the case in many LMICs [1,9]. Despite the proliferation of mobile phones, the Nigerian health sector has not leveraged the advances in mobile technology for MCH delivery, unlike the health sectors in some other LMICs [15]. Similarly, the dominance of mobile apps in the financial and transportation sectors has not translated into the uptake of mobile health apps or telemedicine in the health sector [16]. So far, only a few hospitals in Nigeria have implemented an EHR system in some form [2,7]. However, there is a substantial use of EHRs for programs specific to diseases such as tuberculosis and HIV [17-21].

The challenges of health IT implementation in LMICs, especially Nigeria, include inadequate infrastructure, limited human capacity, brain drain, lack of enforcement of legislation and policies (political will), insufficient financial investment or incentives, and corruption-riddled systems [9,22-24]. Despite funding from the World Health Organization (WHO) and other funding agencies, the implementation is fraught with corruption. Private individuals and organizations in the health system divert the funds earmarked for these IT projects [25]. As a result of these acts, the patients or citizens who are beneficiaries do not get the intended quality of care and health outcomes [24]. Hence, funding agencies should include in funding applications a rider concerning how implementers monitor and evaluate the actual use and effect of resources provided. A very effective tool to achieve this is the development of a theory of change (ToC).

ToC Fundamentals

The origins of the ToC can be traced to Chen and Rossi [26] and Weiss [27], who carried out extensive work in the area of theory-driven and theory-based evaluation. In particular, Weiss [27] popularized the term and modestly defined a "theory of change" as a theory of how and why an initiative will work. This definition seems simplistic; yet, it is foundational. ToC has evolved over the years, considering the ever-changing complexities in international development programs. In this study, we adopt the definition of ToC by the United Kingdom's Department for International Development as "an outcomes-based approach which applies critical thinking to the design, implementation and evaluation of initiatives and programs intended to support change in their contexts" [28]. This definition relates to this feasibility study because this study aimed to bring about change by introducing EHR implementation in a primary health care context.

The ToC is both a process and a product [28-31]. The ToC process articulates the mechanisms of change. The process involves stakeholders who set a long-term goal and go in a reverse direction to specify assumptions and identify preconditions to achieve the desired outcomes [29]. This process leads to the product (ToC map) and is usually developed in versions before, during, and after program implementation. Although there is no single way to design ToCs, it can be asserted that good-quality ToCs should entail certain components such as long-term goals, assumptions, interventions, measurable outcomes, inputs, and outputs [32]. For a ToC to be deemed effective for any program or study such as this EHR

implementation, it should fulfill these 3 criteria: it should be *credible*, *doable*, and *testable* [33]. The combination of assumptions from practitioners' experiences, evidence from literature, findings from previous implementations, and program designer's implicit logic substantiate the credibility of a ToC. In particular, articulating explicit assumptions about the feasibility of the EHR implementation helps to expose, test, and correct the program design logic. The assumptions are like theories that guide each ToC component and their interrelationships, and there is no one-size-fits-all set of assumptions. The assumptions vary from context to context and intervention to intervention [27,34]. On the basis of the specified assumptions, the activities carried out around the intervention will result in outputs, leading to indicators that can be measured to gain the confidence of relevant stakeholders: government, funders and nonprofits, health care workers, and ultimately patients [32].

ToC and Other Relevant Frameworks

There are numerous frameworks used in health informatics, such as the logical framework (logframe) [35], DeLone and McLean (D&M) information systems (IS) success model [36], and examples presented in the WHO digital health monitoring and evaluation guide [37]. These frameworks have a broad purpose of assessing the maturity of an intervention over time but focus on specific criteria or dimensions; for instance, logframes involve logical designing, monitoring, and evaluating inputs, activities, outputs, outcomes, and impacts to achieve the desired results [38]. The logframe approach is very similar to the ToC approach in several ways. Logframes are useful and more linear [30]. Because of the complexity of the EHR intervention and the Nigerian environment, we found that ToCs were more adaptable with regard to capturing the ensuing complex interactions. The D&M IS success model measures the "complex-dependent variable" in IS studies [36]. This model is widely used to assess the interrelationship between critical evaluation dimensions of IT interventions, including information quality, system quality, service quality, system use or use intentions, user satisfaction, and net system benefits [5,39]. In the context of LMICs, the D&M IS success model has been validated by studying electronic hospital IS at 5 Nigerian teaching hospitals [39]. The WHO digital health guide is not a single framework; it examines several evaluation frameworks and illustrates how they could be practically used to support the implementation of digital health interventions in various contexts [37]. Having considered better-known evaluation frameworks, it is worth noting that the ToC scope goes beyond evaluation and covers planning, co-design, stakeholder engagement, and the linkage of causal pathways to individual outcomes. We used the ToC in this study to understand the problem as well as design and evaluate the intervention. The ToC applies to the whole life cycle of the intervention from the creation right through to the evaluation.

Objectives

This study aimed to develop a ToC for the implementation of EHRs for MCH delivery in LMICs. The ToC approach will guide the entire transformation process from paper documentation to EHRs in the study context.

Methods

Setting

The study was conducted at the Festac PHC in Lagos, Nigeria, which has the highest number of physicians (7) and a wider range of health personnel than any other public PHC in Lagos State [40]. With the number of health care staff, the services provided, and operation hours (24 hours, 7 days a week), Festac PHC is a flagship public primary care center in Lagos State known for its role in reducing maternal and child health mortalities [41]. At this facility, patient information was written on paper and maintained in folders and health registers, which posed the issues of confidentiality, missing records, and inefficiencies. As of August 2019, Festac PHC employed 36 health care professionals (HCPs), who served an estimated population of 27,273 residents. There were additional HCPs at the other 16 private clinics and hospitals that serve the same population [40]. A research team funded by the Global Challenges Research Fund [42] through the University of Portsmouth worked with Festac PHC management to conduct a feasibility study for EHR implementation at the health facility. The health facility comprised 6 service departments, including the mother and child center, health records, consultation, general outpatient, laboratory, and pharmacy. At the mother and child center, midwives deliver MCH services and keep patient records in registers meant for services such as ANC, immunization, delivery, and family planning. At the health records unit, health information officers collect and maintain patient information with the help of registers, folders, and filing cabinets. The consultation unit consists of physicians (medical officers) who diagnose patients and keep patients' clinical notes. In the general outpatient department, community health workers (nurses) observe and record patients' vital signs. In the laboratory unit, a laboratory scientist and technicians run tests and maintain test data (specimen source, request, and results) of patients, aiding physicians and midwives in making diagnostic decisions. The pharmacy department consists of a lead pharmacist and pharmacy technicians who order, maintain, and dispense medicines. For this study, 14 participants (n=3, 21% physicians; n=5, 36% midwives and nurses; and n=6, 43% health records officers) were selected using purposive sampling because they were directly involved with patient data at Festac PHC [43]. The study commenced by conducting a remote scoping study in April 2019, which included readiness assessment (through an open-ended interview with the Festac PHC contact person), initial workflow analysis, and risk analysis through email or Skype consultation with the management team of Festac PHC.

Design

We used the ToC approach throughout the life cycle of the implementation to guide the pilot study and identify the preconditions needed to realize the long-term goal of the study [28,30]. Modifications were made from the initial version of the ToC to the revised version to reflect the realities of the implementation process. Because of the complex nature of EHR implementation, we developed and revised ToC maps with the relevant components. The research team developed the first ToC map (Figure 1) as an actual ToC based on evidence from literature, consultation with the local health information

manager, and findings from previous EHR implementations. The ToC map illustrated the *problems* we were trying to solve, the *keystakeholders*, *assumptions*, *inputs*, *intervention*, *outputs*, *measurable effects*, and *wider benefits* of the implementation to realize the *long-term change* [44]. We developed a revised ToC map (Figure 2) to accommodate changes during and after the EHR implementation. These changes related to most of the ToC components and are documented under each component subheading in the Results section. We recognize that implementers should pay attention to sociotechnical issues, especially the interplay between patients' realities and HCPs' mental models and how these influence the EHR design and are represented within the system [45,46].

In the context of this study, the ToC components use these definitions:

- *Long-term change*: the desired goal the stakeholders want to achieve
- *Problems*: the challenges facing the current paper-based health records workflow as highlighted by the stakeholders
- *Stakeholders*: the people directly or indirectly involved or affected by the success or failure of the EHR implementation
- *Assumptions*: the beliefs that specify the underlying reasons for the logical connections that exist among the ToC elements. These beliefs are usually informed by research evidence, clinical practice, and the environment in which the change is taking place.
- *Inputs*: the activities or tasks carried out around the intervention
- *Interventions*: the initiatives or programs embarked on to influence the desired outcomes
- *Outputs*: the tangibles resulting from the inputs and the intervention
- *Measurable effects*: the immediate indicators that can be traced to the implementation process and are readily usable for evaluation. These measures can be quantitative or qualitative.
- *Wider benefits*: generalizable pointers that can guide the stakeholders with regard to the chances of implementing long-term change

The ToC approach is not immune to problems when used as an evaluation tool. Problems of theorizing, measurement, testing, and interpretation are not unusual [27]. To ensure rigor and evaluate the maturity of the implementation, we adapted the success criteria used in the studies by Deriel et al [12] and Fritz et al [47] to supplement the ToC approach. Textbox 1 outlines the categories considered for the success criteria of the implementation and provides definitions for each category.

We engaged the health practitioners and decision-makers at Festac PHC in designing, implementing, and evaluating the EHR system. In particular, the health practitioners at Festac PHC joined in developing the ToC versions, especially providing practical experiences that shaped the theories underpinning the ToC versions. This approach facilitated realistic interactions with the stakeholders and gave a proper understanding of the local context in which the study was conducted [48,49]. We had stakeholder meetings involving the heads of department and EHR champions at the PHC at the start and during the implementation process. Each stakeholder discussed the issues of the existing paper-based health record system and their expectations and experiences of the new EHR system, which validated the findings of the first ToC map. Subsequently, health informatics experts validated the revised ToC findings at the MedInfo 2019 conference in Lyon, France.

We developed a generic version of the ToC map (Figure 3) to reflect a holistic framework as a toolkit for relevant stakeholders who want to embark on this kind of intervention in similar contexts beyond Lagos, Nigeria. The stakeholders can adapt it for EHR implementations in primary care settings but need to pay close attention to inherent characteristics in these environments. Despite the nuances in different contexts, the process and steps involved in the creation of the ToC map are not to be ignored. Chen and Rossi [26] stressed the importance of giving adequate attention to understanding the implementation process and not being too concerned about whether the initiative has yielded excellent results.

Figure 1. An initial version of the theory of change for the scheduled electronic health record (EHR) implementation at Festac Primary Health Centre (PHC). ANC: antenatal care; FHIR: Fast Healthcare Interoperability Resources; GCRF: Global Challenges Research Fund; OpenMRS: Open Medical Records System; UoP: University of Portsmouth.

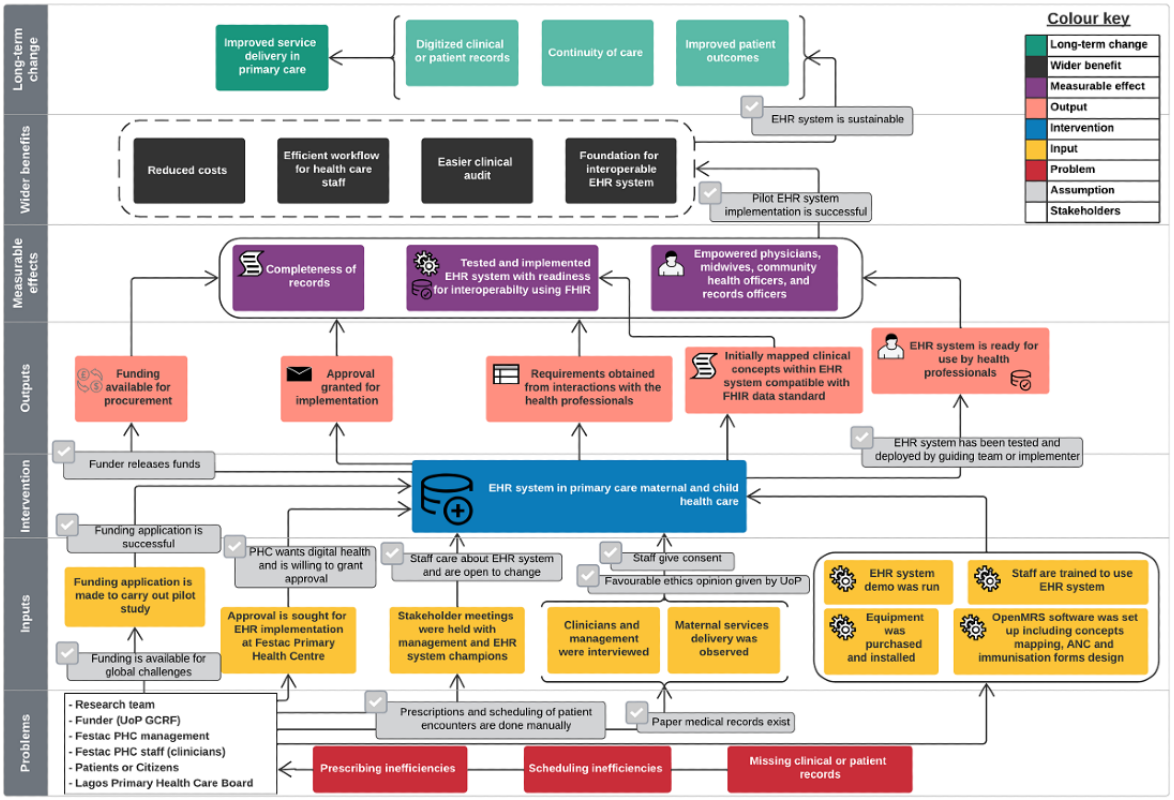
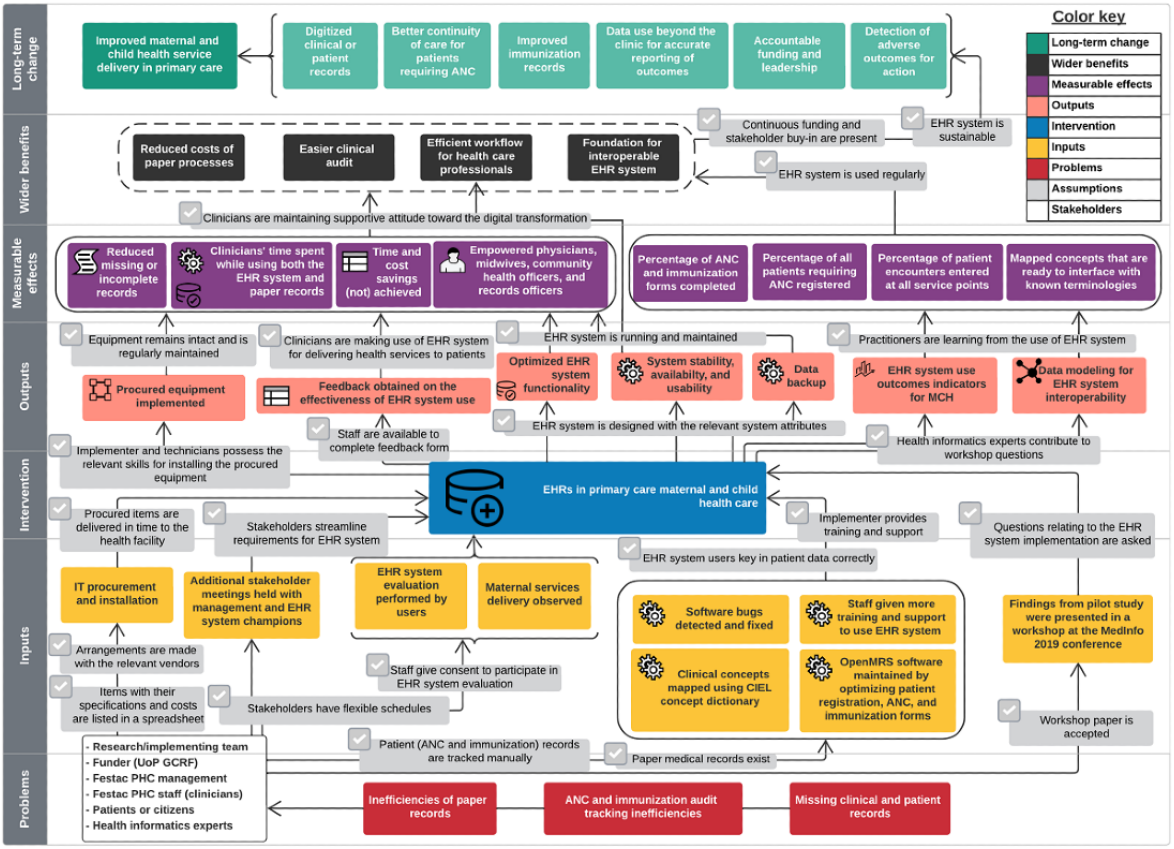


Figure 2. A revised version of the theory of change for electronic health record (EHR) implementation at Festac Primary Health Centre (PHC), including findings from a workshop at the MedInfo 2019 conference. ANC: antenatal care; CIEL: Columbia International eHealth Laboratory; GCRF: Global Challenges Research Fund; MCH: maternal and child health care; OpenMRS: Open Medical Records System; UoP: University of Portsmouth.

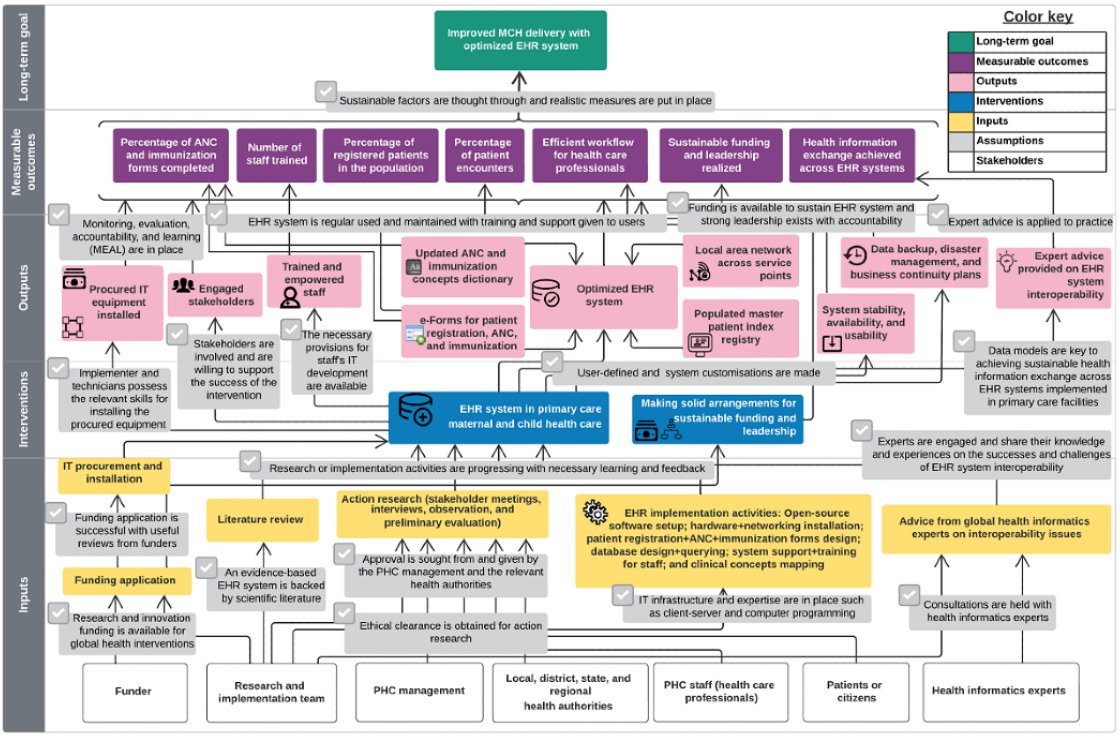


Textbox 1. Categories for success criteria and their definitions for electronic health record implementation (adapted from Deriel et al [12] and Fritz et al [47]).

Categories and definitions

- Ethics
  - Regulatory and cultural issues such as health data security, privacy, and confidentiality
- Political
  - Health policies and countrywide circumstances, including health care infrastructure, characteristics, ministries of health, and primary health care boards
- Organizational
  - Managerial circumstances within the organization itself, including human resources, skilled staff, or local buy-in; leadership and governance; project management and commitment to implementation; and data use
- Financial
  - Resources (including human and equipment) and funding
- Functionality
  - System features and functions, including modules, data handling, forms, and reports
- Technical
  - Infrastructure, software architecture, user interfaces, data standards, and privacy or security
- Training
  - Skills training as well as computer literacy and educational background and user support
- Sustainability
  - Transition from external stakeholder to local management across all categories, including financing

Figure 3. A generic version of the theory of change for electronic health record (EHR) implementation, without context-specific details. ANC: antenatal care; MCH: maternal and child health care; PHC: primary health center.



EHR System Selection

Open Medical Records System (OpenMRS) is an EHR software program built for low-resource settings to improve health care delivery with the help of a global community that supports the software [50]. We selected OpenMRS as the EHR application for the pilot implementation because it is an open-source program and therefore freely available, which fits into the funding realities of LMICs, including Nigeria. The OpenMRS software source code can be modified and tailored to the needs of the particular context in which it is being used. It is an enterprise platform with flexible modules that have matured over time and been implemented in similar settings with a vibrant web-based community of developers and implementers [51,52]. We adapted existing OpenMRS modules to facilitate the identified use cases such as patient registration, outpatient clinic, laboratory, and mother and child clinic to manage clinical workflows. Moreover, we adapted UgandaEMR's ANC and immunization e-forms to save development time and initial user-testing requirements.

Ethics Approval

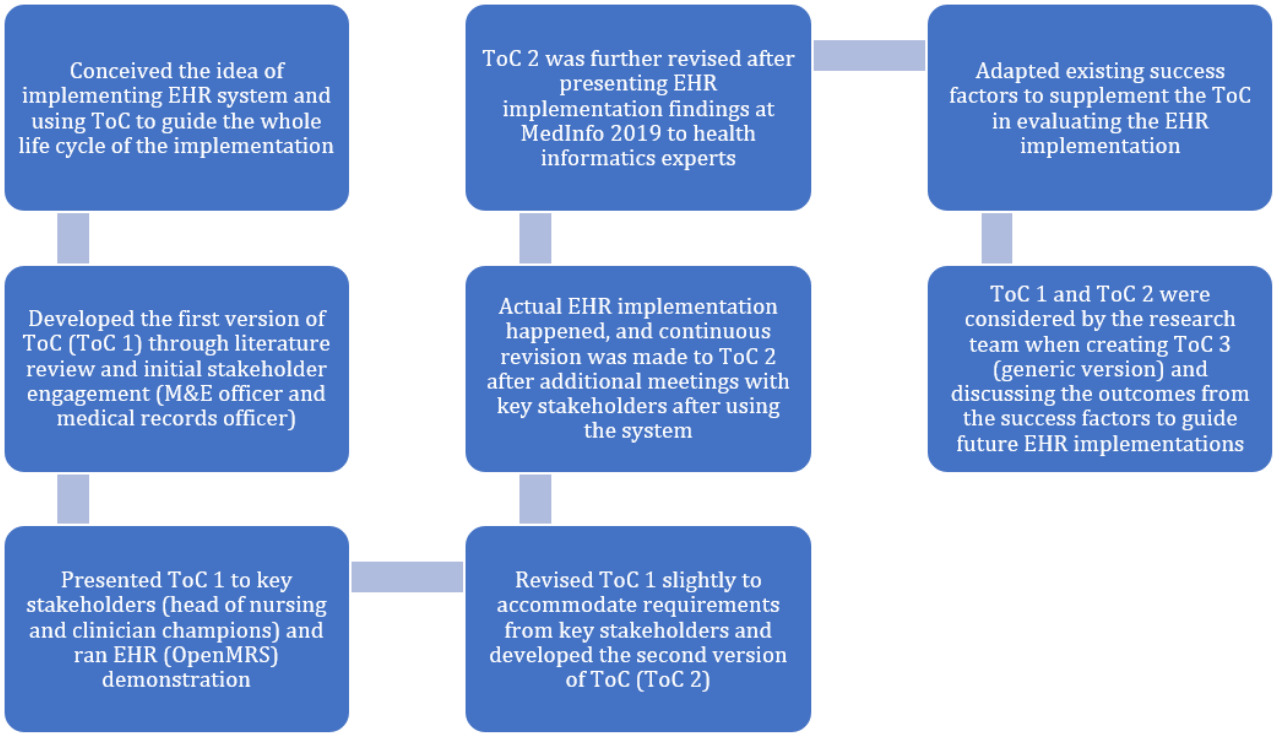
This study obtained a favorable opinion from the University of Portsmouth Faculty of Technology ethics committee (TECH2019-T.A-01). Participation in the study was voluntary, and participants were free to withdraw at any time without giving any reason. The participants provided written consent by completing a participant consent form. The study considered the security, privacy, and confidentiality of patient records from the outset. The paper health records were kept locked in a card room at the PHC. Although the reception area is positioned close to the card room, at busy times anyone could access the room with malicious intentions to cart away or damage the paper records. Hence, the EHR implementation took into account secure access to the electronic records by creating user accounts for relevant clinicians, ensuring that only users authorized by the heads of department could access the system [4].

Results

Overview

In this section, we report the complete ToC life cycle (Figure 4) for this study commencing from idea conception to the development of the initial ToC map and revised ToC map, illustrating how we accomplished the EHR implementation tasks at Festac PHC. At the same time, we hypothesize that program designers and relevant stakeholders can adapt the generic ToC map for EHR implementations in similar contexts. Subsequently, we provide a detailed narrative of the long-term change and identified preconditions from the ToC process. From this process, we produced a summary of the key successes and lessons learned alongside the study's implications to evaluate the process (Multimedia Appendix 1).

Figure 4. Complete theory of change (ToC) life cycle for electronic health record (EHR) implementation at Festac Primary Health Centre. M&E: monitoring and evaluation; OpenMRS: Open Medical Records System.



ToC Life Cycle

Figure 4 illustrates the entire ToC process for the EHR implementation and the key changes that occurred along the way. The ToC process is important because it helps to identify all the key stakeholders; for example, it helped to identify the significance of having clinical stakeholders evaluate the initial ToC. The conversations with the key stakeholders influenced the revised version of the ToC. Moreover, the process helped to identify problems early as well as the changes in direction for the EHR implementation, saving time, and cost.

Long-term Change

Initially, the desired goal of the study was to achieve improved service delivery in primary care with the use of an EHR system. The allied goals were digitization of patient records, continuity of care, and improved patient outcomes. However, the overarching goal was slightly modified to accommodate practitioners' assumptions. Hence, the long-term change is to achieve improved MCH service delivery in primary care using the EHR system. At Festac PHC, along with this long-term goal, the goals are digitization of patient records, better continuity of care for women using the ANC service, and improved handling of children's immunization records. We theorize, with the generic ToC map, that the long-term goal is to improve MCH service delivery with an optimized EHR system based on the assumption that sustainability factors have been thought through, and measures have been put in place to achieve this goal.

Assumptions

The initial ToC map served as the basis for the actual implementation, with preliminary assumptions emanating from the prior knowledge of the research team, the literature of existing EHR implementations, and initial conversations with the monitoring and evaluation officer and a medical records officer. The initial assumptions included the following:

1. Paper medical records exist.
2. Prescriptions and scheduling of patient encounters are carried out manually.
3. Funding is available for global challenges.
4. Funding application is successful.
5. PHC wants a digital health program and is willing to grant approval.
6. Favorable ethics opinion is given by the University of Portsmouth.
7. Funder releases funds for procurement of equipment.
8. Stakeholders care about EHRs and are open to change.
9. Practitioners give consent to be interviewed and observed at the health facility.
10. EHR system is tested and deployed by the guiding team and implementer.
11. Pilot EHR implementation is successful.
12. EHR system is sustainable.

After the actual EHR implementation, the initial ToC map was revised to reflect the real changes encountered during the pilot study; for example, priorities for the EHR system shifted from scheduling and prescription to ANC and immunization. At the time of developing the initial ToC map, the Festac PHC stakeholders had identified the need for booking patient appointments and producing prescriptions electronically with the EHR system. However, after the face-to-face stakeholder meeting at the health facility, the practitioners noted that e-forms for ANC and immunization were their immediate needs for the EHR system. Another change to the ToC revision was the shift in networking design from the cloud to a local area network. This shift was due to connectivity problems and a lack of guarantees from the management regarding sustaining the internet subscription payment. This is the dominant approach to EHRs in LMICs because few of the smaller sites can guarantee reliable internet connectivity for cloud-based use, although certain LMICs do this well [14,53].

In addition, the revised ToC included findings from the research workshop (MedInfo 2019 conference), where the EHR use outcomes from the pilot study were presented. Global health informatics experts offered advice at the workshop, during which it was emphasized that data models are key to realizing effective communication exchange across digital health systems by adopting the appropriate interoperability standards for MCH, well-known examples of which are Fast Healthcare Interoperability Resources [54] and OpenEHR [55]. In addition, the drivers for an interoperable EHR system differ between LMICs and HICs; for example, LMICs focus mainly on aggregate data from the health information system for disease control, population health monitoring, and health policy and planning. Funders use these aggregate data to drive health financing and, in some cases, to fund EHR implementations. However, HICs pay more attention to the quality of care, continuity of care, and precision medicine. In addition, adequate infrastructure and accountable funding were identified to be key preconditions needed for a sustainable EHR implementation. In sum, toolkits are important in shaping EHR implementations for MCH services.

Although some *assumptions* stay the same, others were modified. Textbox 2 illustrates these assumptions and how they were generated.

For the improvement of MCH services to be achieved, it was assumed that the EHR system was sustainable. The EHR system needs to be used regularly to bring about the broader benefits of its implementation.

Textbox 2. Assumptions and their sources.

Assumptions and sources

- Antenatal care and immunization records are tracked manually
  - Practitioners
- Items with their specifications and costs are listed in a spreadsheet
  - Electronic health record implementation
- Arrangements are made with the relevant vendors
  - Program designer
- Procured items are delivered in time to the health facility
  - Electronic health record implementation
- Implementer and technicians possess the relevant skills for installing procured equipment
  - Policy makers and program designer
- Stakeholders have flexible schedules
  - Practitioners
- Staff give consent to participate in the electronic health record system evaluation
  - Practitioners
- Stakeholders streamline requirements for electronic health record system
  - Practitioners and program designer
- Implementer provides training and support
  - Policy makers, practitioners, and program designer
- Electronic health record users key in patient data correctly
  - Practitioners and program designer
- Workshop paper is accepted
  - Program designer
- Questions relating to the electronic health record implementation are asked by workshop participants
  - Health informatics experts and program designer
- Health informatics experts contribute to workshop questions
  - Health informatics experts and program designer
- Staff are available to complete a feedback form
  - Practitioners
- Electronic health record system is designed with the relevant system attributes
  - Practitioners and program designer
- Hardware equipment and electronic health record system software remain intact and are maintained regularly
  - Policy makers, practitioners, and program designer
- Clinicians are making use of electronic health records regularly for delivering health services to patients
  - Practitioners and policy makers
- Stakeholders are learning from electronic health record use and data

- Policy makers, practitioners, and program designer
- Clinicians are maintaining a supportive attitude toward digital transformation
  - Policy makers, practitioners, and program designer

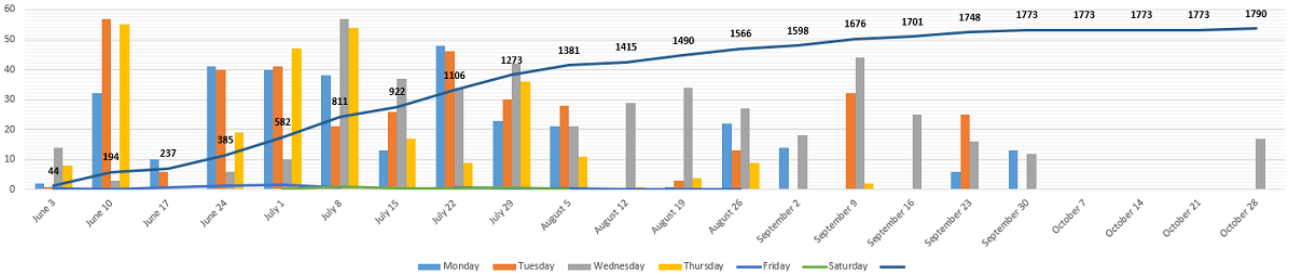
Wider Benefits

On the basis of the assumption that the pilot EHR implementation is successful, there would be benefits accrued to Festac PHC. These benefits include reduced costs of paper processes, including expenses for stationery; efficient workflow for the health care staff; easier clinical audit of patient records; and readiness for a sustainable EHR system. The sustainability of the EHR system will enable effective health information exchange as the use of EHRs becomes widespread over time.

Measurable Effects

It was anticipated that the availability of the EHR system alongside the surrounding outputs would result in the completeness of health records, which could be measured against the use of paper health records by the health practitioners. Another potentially measurable effect concerns clinician time spent using both paper and EHR systems [56]. During the implementation, we found that clinicians spent more time using both paper and electronic systems simultaneously, which affected the EHR system's complete records outcome. We enrolled 14 clinicians to use the EHR system, and Figure 5 shows the rate of EHR system adoption and use for the study's first phase lasting for 5 months (June 2019 to October 2019) from the time the system went live. A total of 2799 encounter forms were completed on the EHR system; 1790 (63.95%) patients were registered, with an equivalent number of patient registration forms being completed. ANC and immunization encounter forms (198/2799, 7.07% and 309/2799, 11.04%, respectively) were completed. Vital signs (325/2799, 11.61%) and visit notes (177/2799, 6.32%) were entered into the EHR system. Of the 325 vital signs forms, 148 (45.5%) consultations were not recorded using the visit notes because some physicians only used paper notes. A major system downtime occurred from October 3 to 29, 2019, which affected data entry. Longer-term success factors, which are yet to be measured, are the realization of funding sustainability and accountable leadership, as well as health information exchange achieved between the EHR system and other health IS.

Figure 5. Cumulative daily and weekly data entry with regard to patients registered on the electronic health record system.



Outputs

The study received a letter of approval for the implementation from the local authority. This approval enabled the release of funds and the travel of a research team member (TA) to the health facility. The funder released the funds to procure the IT equipment needed for the study. The interactions with the health practitioners made it possible to obtain the requirements to design and develop the intervention. After we incorporated the active inputs of various stakeholders, the EHR system was ready for use by the health practitioners. The outputs in the revised ToC map were procured equipment, feedback from EHR use, optimized EHR system functionality, EHR use outcomes indicators for MCH, and data modeling for EHR system interoperability. Other key outputs were the critical system attributes (such as system stability, availability, and usability) and full and incremental data backup of patient records to the cloud. In the event of system damage, fire, flooding, or any adverse incidents, the PHC can restore the records from the backup.

Intervention

The main intervention for this study was the introduction of an EHR system in primary care MCH services. Initially, problems were perceived based on explicit and implicit assumptions about paper medical records and prescriptions and scheduling of patient encounters being carried out manually. After face-to-face stakeholder meetings on site, the practitioners were of the unanimous opinion that prescribing and scheduling inefficiencies were not the priority issues; rather, priority should be accorded to paper records handling, ANC and immunization-tracking inefficiencies, and missing patient records. These problems validated the introduction of the EHR intervention at Festac PHC.

Stakeholders

The stakeholders are the research team (TA, PS, and HF), funder, Festac PHC management (local authority), primary health care board, health care practitioners, patients, and health informatics experts. They carried out several activities at various stages of the study. The research team made some informal contacts with the local primary care facility authorities to understand their problems and the desired long-term outcomes.

The team reviewed existing studies to gain background knowledge of previous EHR implementations in similar contexts. After the review, the team developed the initial version of the ToC map, informed by the explicit assumptions of the practitioners and implicit assumptions gleaned from previous implementations. The research team prepared a funding application and sought approval for the pilot study. Both the funder and the local health authority approved the pilot study.

Technical Implementation

There was a demonstration of the OpenMRS software during the first stakeholder meeting. The activity helped the practitioners to have a feel of how the intervention works. Before this meeting, the contact person from the PHC had been testing the demonstration version of the EHR system; they gave feedback on what the PHC specifically wanted. The main technical components of OpenMRS are the database (eg, data concepts mapping, backups, and security) and the EHR software (clinical modules and customizations). The research team initially designed a cloud solution before the implementation but changed to a local area network design because of poor internet access at the health facility. Through a combination of on-site and remote support, the research team contributed to installing and configuring the software. The equipment included laptops, a desktop PC (dedicated server), networking equipment (16-port Ethernet switch, wireless router, Category 6 cables, and RJ45 connectors), a power inverter (to provide power for the server when electricity from the national grid and generator set is unavailable), and a printer.

Discussion

Principal Findings

This study shows the value of the ToC process for robust planning, analysis, and evaluation of EHR implementation complexities, as well as challenging the assumptions of all stakeholders. The process requires logical reasoning, effectively engaging stakeholders in drawing implicit assumptions, designing the preconditions, and mapping the ToC backward from the long-term goal to inputs. Political factors play a role in influencing what practitioners say about their beliefs or theories regarding the desired change. The practitioners may have concerns about the management's disapproval of their assumptions [33]; for example, we asked the HCPs about the leadership style of their line managers and the effect it has on their use of the EHR system. Some (7/14, 50%) of the HCPs made positive comments about their managers. Although it is possible to have all-positive feedback about leadership styles in a typical work setting, the lack of concerns or negative comments may suggest desirability bias or groupthink [57].

A ToC is useful in articulating assumptions made about a program or intervention to achieve its desired results. We generated assumptions from peer-reviewed evidence (documents and prior research); experience and views of practitioners and other stakeholders such as funder, government, and policy makers; and logical reasoning (Textbox 2). However, it can be problematic to test assumptions even when they are explicitly stated. Problems such as measurement, generalization, and validation usually plague program theory [27]. Our study extensively evaluated the ToC-based implementation using previously defined success criteria across multiple dimensions of implementation and use (Multimedia Appendix 1) [12,47], which is a methodological innovation in LMIC settings because of the wide range of evaluation criteria. However, combinations of some individual criteria have been used. Certain authors have argued that theory-based evaluation such as the ToC is more a methodology than a theory because it uses different research methods (eg, randomized controlled trials, interviews, and workshops) for its development [30,44]. Weiss [33] argues that the ToC is an approach and a theory because it is built on assumptions (beliefs), preconditions, inputs, and outputs, which influence the way people behave.

Again, the ToC approach is particularly useful in capturing the complexities of a program relating to its outcomes, outputs, inputs, and activities to bring about long-term change by using relevant interventions [58]. The research team engaged the relevant stakeholders by asking them to share their experiences and practices (explicit assumptions). We drew out the implicit assumptions, which were not obvious to the practitioners and experts, through interviews and a workshop (findings to be published), and then modeled these assumptions and combined them with evidence and logic, all of which were put together in readiness to transfer into practice.

Reflections Based on Experiences of EHR Implementations in Other LMICs

Despite Festac PHC being an early adopter of the EHR system and the only one among public PHCs in Lagos State, the management has not done enough in terms of funding the infrastructure and ensuring its sustainability. The issue of funding and other EHR implementation challenges are not peculiar to the Nigerian context; rather, they are applicable to different LMIC contexts [51,53]. Comparison evaluations of EHR systems in LMICs were provided by 2 papers, published in 2017 and 2018 (Multimedia Appendix 2 [51,53]). Although there is anecdotal evidence of EHR implementations across Nigeria, there is no known peer-reviewed evidence of OpenMRS implementation in the country. As of June 2021, the OpenMRS HIV Reference Implementation initiative funded by the Centers for Disease Control and Prevention is supporting >1000 site rollouts of OpenMRS in Nigeria as well as improvements in the user interface, reporting, and other initiatives [59]. A recent paper [60] tried to examine the impact of OpenMRS implementations globally over a 15-year period, but no concrete evidence on Nigeria was available, except for some brief mentions. This study should help to address this gap, especially where public primary care in Nigeria is concerned.

Multimedia Appendix 2 compares findings from OpenMRS implementations in 3 LMICs (Nigeria, Sierra Leone, and Kenya), inclusive of this study (Festac PHC in Nigeria). Common findings across the 3 studies related to data collection, staff training, and infrastructure. These studies showed that EHR use results in clinical workflow efficiencies. At the same time, the studies discussed the challenges encountered during implementation, which centered mainly on inadequate infrastructure, funding, dedicated IT support, and stakeholder buy-in. A significant issue across the 3 EHR implementations

is sustainability, and our Nigerian (Festac PHC) study used the ToC approach to underscore this issue extensively. Despite their successful completion, the implementations did not continue beyond the first or second phase. Hence, stakeholders must pay close attention to sustainability issues before embarking on EHR implementations in LMICs.

Reflections Based on Experiences of EHR Implementations in HICs

Policy makers and politicians in LMICs can learn from countries that incentivized EHR adoption by providing implementation funds to health facilities. A prime example is the United Kingdom, where the EHR adoption rate in primary care, particularly general practitioner (GP) practices, is nearly 100% [61,62]. Among other factors, financial incentives from the government have proven to be an effective impetus for EHR implementation across GP practices. For many years, thought leaders in the GP profession have collaborated with the government to provide incentives for digitizing practices and eliminating barriers. Hence, GPs were more willing to use EHRs than hospital physicians, helping the former leverage the successful health IT intervention [62]. However, despite the successful EHR adoption rate by GP practices in the United Kingdom, the system has its shortcomings: it sometimes fails as patients show up at the community pharmacy expecting to pick up their medications only to find that the electronic prescription has not reflected in the pharmacy system. This issue can often delay treatment for patients, especially on weekends when GP practices are closed, and the pharmacy team chases prescriptions. The GP's on-call team can usually access the system and fax the prescriptions to the pharmacy, but the effectiveness of this process varies across the United Kingdom.

The US government program based upon the Health Information Technology for Economic and Clinical Health Act of 2009 provides financial incentives to physician practices and hospitals to foster digital health implementation and improve the quality of care for patients. These incentives have since led to the widespread adoption and meaningful use of EHR systems across all levels of health care in the United States, with the resultant digital health transformation and improved clinical outcomes [63-65]. However, rapid implementation of existing EHR systems has been associated with many challenges in workflow, usability and physician stress or overload. The UK model of adoption of primary care EHR systems may be better in terms of a limited number of carefully vetted systems, low costs, and robust interoperability with many hospitals; for example, in West Yorkshire [66].

Reflections on Data Entry at Festac PHC

Inconsistencies in EHR data entry during patient encounters occur because of several factors, including human, organizational, and system factors. The willingness of clinical staff to use the new system was lacking because of the perception that the system would add to their existing workload, reflecting the realities of data entry operations and the shortage of health workers in LMICs [67]. Only a few HCPs were keen on using the system. Hence, little or no data entry is completed if the active HCPs are not on duty. Sometimes, the HCPs attend staff verification exercises, leaving the EHR system in the hands of casual staff who do not have permission to use it because of clinical accountability requirements. Lack of leadership motivation or incentive to use the system could prevent health information officers, physicians, nurses, and midwives from understanding the need to work on data entry. System downtime happens occasionally; when this happens, there is no health IT support technician on the ground to resolve the issue, and hence the PHC relies on the implementer, who, although not contractually obliged, may sometimes help out. To resolve system issues, the PHC management could employ an IT support technician on a full-time or part-time basis, but the management should be keen and be ready to include the employment cost in the clinic's budget. In a recent review on the importance of primary care records in LMICs, we found that there seems to be a particular challenge with EHR data collection in primary care organizations [68]; for example, MCH EHR data collection was challenging because of local factors such as the level of technology available for data entry at the point of childbirth. Hence, this is a larger problem for people who run modest primary care EHR systems in LMIC settings, a problem not specific to Nigeria. This implementation study successfully demonstrated improvements in MCH services data collection. However, the lack of effective human, organizational, and system support is responsible for inconsistent data entry in the EHR system, leading to poor clinical benefits and inaccurate reporting.

The ToC approach gave insights into the potential causes of the breakdown of the system, such as the issues concerning regular use and data entry by key staff, which allowed for provision of additional planning and training. A simple *cost-benefit* approach to framing the overall implementation process to determine the likely gains (value) to staff, patients, health systems, and funders would be helpful. It would be valuable to determine whether these costs outweigh the challenges of learning to use the system and the pain of working on data entry. In addition, the proposed investment in infrastructure and support could be balanced by the concrete benefits. The costs often fall on staff working on data entry who do not benefit much from the outputs. Hence, the combined effect of the utility of an application and ease of use gives stronger predictability for actual use, which is incorporated in the D&M model.

There is a growing interest in alternative data entry approaches, including the "scribe" model (in US primary care) [69], natural language processing–enabled data capture, and optical mark recognition (OMR). These alternative approaches could address the issue of clinicians' avoidance of using the EHR system. The "scribe" model introduces a way of working where a human scribe (a volunteer or health professional) manually enters the applicable information such as observations, diagnosis, and test results into the EHR during the patient visit as spoken aloud by the physician or nurse [70]. However, this could affect clinical data quality because the scribe might not be a suitably qualified clinician and prone to making data entry errors, which, in turn, could affect health outcomes. Natural language processing data capture applications allow HCPs, especially physicians, to capture structured data with unstructured dictation into the EHR [71]. OMR is a nondictation, scanning method of data capture where the OMR software processes paper clinical forms that

have been scanned with a modest office scanner or low-cost document camera [72]. This approach ensures that clinicians who record clinical data on paper do not also have to enter the data once or twice in other records. It requires stability of systems, a person to oversee the scanning and data extraction, and user confidence. It might develop as a model to overcome a data entry backlog in the EHR system, increasing the value for clinicians, particularly if recent improvements in optical character recognition software can be shown to be effective in interpreting structured handwriting.

Limitations

This study includes several limitations with regard to developing the ToC. First, the research team was extensively involved in developing and revising the ToC map, which may have contributed to a social desirability bias. Second, the first author (TA) mainly worked on the analysis of the ToC maps under the guidance of the last author (PS) and the second author (HF). We would have engaged the HCPs and stakeholders in the analysis, but they were not well versed with the technicalities of the ToC approach. Future studies will ensure that HCPs are familiarized with the ToC analysis. The relevant stakeholders were fully engaged in the clinical, data collection (interviews and observations), and managerial aspects of the design.

Conclusions

This research presented the ToC as a rewarding approach to framing dialogue with stakeholders. It functioned as a valuable framework for planning an EHR implementation and the steps needed to define the requirements and success factors, likelihood of longer-term success, and evaluation metrics. For new implementers, knowing how to structure this implementation process could be very useful. Future health IT implementation in primary care can adapt the ToC approach to their contexts with necessary modifications based on inherent characteristics. The pilot EHR implementation served as a small-scale foundation that can support health information exchange and as a digital health exemplar for other PHCs in Lagos State and Nigeria. Other health care providers can learn from, and build on, the implementation to support the delivery of MCH and other health services. Furthermore, the pilot EHR system represented a digital enabler that provides computable and machine-readable health data, the necessary first step toward more complex aspects such as interoperability, clinical decision support, and a learning health system. Further work is needed to extend the scope of the implementation to cover other public PHCs. There is a need to secure more funds for additional infrastructure alongside solid leadership to ensure sustainability and scalability. In addition, it will be helpful to explore the interoperability of health data across public PHCs by designing a national health data model for an MCH services data set. The model should be based on established data standards and an examination of the preconditions and drivers for implementing such a model and build on existing work on clinical decision support for MCH services [73].

Acknowledgments

The authors thank the management and staff of Festac Primary Health Centre for granting approval to, and supporting, the study. The pilot study was funded by the Global Challenges Research Fund allocation to the University of Portsmouth.

Authors' Contributions

TA and PS conceived the study. TA drafted the manuscript. All authors revised and approved the final manuscript.

Conflicts of Interest

HF is a cofounder of the Open Medical Records System open-source software project that developed the software used in this study; he contributed to the drafting and revision process.

Multimedia Appendix 1
Summary of successes achieved and lessons learned from the pilot study at Festac Primary Health Centre, as well as implications for electronic health record implementations in Nigeria and other low- and middle-income countries.
[DOCX File , 21 KB - medinform_v10i8e33491_app1.docx]

Multimedia Appendix 2
A comparison of electronic health record implementation findings from 3 studies conducted in low- and middle-income countries.
[DOCX File , 21 KB - medinform_v10i8e33491_app2.docx]

References

1. Rumball-Smith J, Ross K, Bates DW. Late adopters of the electronic health record should move now. BMJ Qual Saf 2020 Mar;29(3):238-240 [FREE Full text] [doi: 10.1136/bmjqs-2019-010002] [Medline: 31732701]
2. Odekunle FF, Odekunle RO, Shankar S. Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region. Int J Health Sci (Qassim) 2017;11(4):59-64 [FREE Full text] [Medline: 29085270]

3. Attah A. Implementing an electronic health record in a Nigerian secondary healthcare facility. Prospects and challenges. UiT Norges arktiske universitet. 2017. URL: https://munin.uit.no/handle/10037/12245 [accessed 2020-03-07]
4. Adamu J, Hamzah R, Rosli MM. Security issues and framework of electronic medical record: a review. Bulletin EEI 2020 Apr 01;9(2):565-572. [doi: 10.11591/eei.v9i2.2064]
5. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: an evaluation of information system impact and contingency factors. Int J Med Inform 2014 Nov;83(11):779-796. [doi: 10.1016/j.ijmedinf.2014.06.011] [Medline: 25085286]
6. Waterson P. Health information technology and sociotechnical systems: a progress report on recent developments within the UK National Health Service (NHS). Appl Ergon 2014 Mar;45(2):150-161. [doi: 10.1016/j.apergo.2013.07.004] [Medline: 23895916]
7. Ebenso B, Allsop MJ, Okusanya B, Akaba G, Tukur J, Okunade K, et al. Impact of using eHealth tools to extend health services to rural areas of Nigeria: protocol for a mixed-method, non-randomised cluster trial. BMJ Open 2018 Oct 18;8(10):e022174 [FREE Full text] [doi: 10.1136/bmjopen-2018-022174] [Medline: 30341123]
8. Reis ZS, Maia TA, Marcolino MS, Becerra-Posada F, Novillo-Ortiz D, Ribeiro AL. Is there evidence of cost benefits of electronic medical records, standards, or interoperability in hospital information systems? Overview of systematic reviews. JMIR Med Inform 2017 Aug 29;5(3):e26 [FREE Full text] [doi: 10.2196/medinform.7400] [Medline: 28851681]
9. Oluoch T, de Keizer NF. Evaluation of health IT in low-income countries. Stud Health Technol Inform 2016;222:324-335. [Medline: 27198114]
10. Culture of quality and safety: a prerequisite for any informatics intervention. In: Global Health Informatics: Principles of eHealth and mHealth to Improve Quality of Care. Cambridge, Massachusetts, United States: The MIT Press; 2017.
11. What if a smart phone could save a life? Virtual Doctors. URL: https://www.virtualdoctors.org/ [accessed 2020-03-07]
12. deRiel E, Puttkammer N, Hyppolite N, Diallo J, Wagner S, Honoré JG, et al. Success factors for implementing and sustaining a mature electronic medical record in a low-resource setting: a case study of iSanté in Haiti. Health Policy Plan 2018 Mar 01;33(2):237-246. [doi: 10.1093/heapol/czx171] [Medline: 29253138]
13. Mirzoev T, Etiaba E, Ebenso B, Uzochukwu B, Manzano A, Onwujekwe O, et al. Study protocol: realist evaluation of effectiveness and sustainability of a community health workers programme in improving maternal and child health in Nigeria. Implement Sci 2016 Jun 07;11(1):83 [FREE Full text] [doi: 10.1186/s13012-016-0443-1] [Medline: 27268006]
14. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2018 Mar 06;25(S 01):S48-S61. [doi: 10.15265/iys-2016-s006]
15. Lee SH, Nurmatov UB, Nwaru BI, Mukherjee M, Grant L, Pagliari C. Effectiveness of mHealth interventions for maternal, newborn and child health in low- and middle-income countries: systematic review and meta-analysis. J Glob Health 2016 Jun;6(1):010401 [FREE Full text] [doi: 10.7189/jogh.06.010401] [Medline: 26649177]
16. Eze E, Gleasure R, Heavin C. Reviewing mHealth in developing countries: a stakeholder perspective. Procedia Comput Sci 2016;100:1024-1032. [doi: 10.1016/j.procs.2016.09.276]
17. Were MC, Nyandiko WM, Huang KT, Slaven JE, Shen C, Tierney WM, et al. Computer-generated reminders and quality of pediatric HIV care in a resource-limited setting. Pediatrics 2013 Mar;131(3):e789-e796 [FREE Full text] [doi: 10.1542/peds.2012-2072] [Medline: 23439898]
18. Oluoch T, Katana A, Kwaro D, Santas X, Langat P, Mwalili S, et al. Effect of a clinical decision support system on early action on immunological treatment failure in patients with HIV in Kenya: a cluster randomised controlled trial. Lancet HIV 2016 Feb;3(2):e76-e84. [doi: 10.1016/s2352-3018(15)00242-8]
19. Amoroso C, Akimana B, Wise B, Fraser H. Using electronic medical records for HIV care in rural Rwanda. In: Studies in Health Technology and Informatics. Amsterdam, Netherlands: IOS Press; 2010.
20. Douglas GP, Gadabu OJ, Joukes S, Mumba S, McKay MV, Ben-Smith A, et al. Using touchscreen electronic medical record systems to support and monitor national scale-up of antiretroviral therapy in Malawi. PLoS Med 2010 Aug 10;7(8):e1000319 [FREE Full text] [doi: 10.1371/journal.pmed.1000319] [Medline: 20711476]
21. Fraser H, Habib A, Goodrich M, Thomas D, Blaya J, Fils-Aime J, et al. E-health systems for management of MDR-TB in resource-poor environments: a decade of experience and recommendations for future work. Stud Health Technol Inform 2013;192:627-631. [Medline: 23920632]
22. National health ICT strategic framework 2015-2020. Federal Ministry of Health. URL: https://www.health.gov.ng/doc/HealthICTStrategicFramework.pdf [accessed 2020-05-27]
23. Implementing the basic health care provision fund in Nigeria: a framework for accountability and good governance internet. Resilient & Responsive Health Systems. URL: https://resyst.lshtm.ac.uk/resources/implementing-the-basic-health-care-provision-fund-in-nigeria-a-framework-for [accessed 2022-07-14]
24. Mackey TK, Vian T, Kohler J. The sustainable development goals as a framework to combat health-sector corruption. Bull World Health Organ 2018 Jun 04;96(9):634-643. [doi: 10.2471/blt.18.209502]
25. Olaronke I, Ishaya G, Rhoda I, Janet O. Interoperability in Nigeria healthcare system: the ways forward. Int J Inf Eng Electron Bus 2013 Oct 01;5(4):16-23. [doi: 10.5815/ijieeb.2013.04.03]
26. Chen H, Rossi PH. Evaluating with sense. Eval Rev 2016 Jul 26;7(3):283-302. [doi: 10.1177/0193841x8300700301]

27. Weiss CH, Connell JP. Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In: New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts. Washington, DC: The Aspen Institute; 1995.
28. Review of the use of 'Theory of Change' in international development. UK Department of International Development. URL: https://www.gov.uk/government/news/dfid-research-review-of-the-use-of-theory-of-change-in-international-development [accessed 2020-02-03]
29. Taplin D, Clark H, Collins E, Colby D. Theory of change TECHNICAL PAPERS a series of papers to support development of theories of change based on practice in the field. ActKnowledge. URL: http://www.theoryofchange.org/wp-content/uploads/toco_library/pdf/ToC-Tech-Papers.pdf [accessed 2020-03-23]
30. De Silva MJ, Breuer E, Lee L, Asher L, Chowdhary N, Lund C, et al. Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. Trials 2014 Jul 05;15:267 [FREE Full text] [doi: 10.1186/1745-6215-15-267] [Medline: 24996765]
31. Daruwalla N, Jaswal S, Fernandes P, Pinto P, Hate K, Ambavkar G, et al. A theory of change for community interventions to prevent domestic violence against women and girls in Mumbai, India. Wellcome Open Res 2019 Mar 25;4:54. [doi: 10.12688/wellcomeopenres.15128.1]
32. Connell JP, Kubisch AC. Applying a theory of change approach to the evaluation of comprehensive community initiatives: progress, prospects, and problems. Educ Crisis Conflict 1998;2(15-44):1-6 [FREE Full text]
33. Weiss CH. How can theory-based evaluation make greater headway? Eval Rev 2016 Jul 26;21(4):501-524. [doi: 10.1177/0193841x9702100405]
34. Nkwake A. Working with Assumptions in International Development Program Evaluation. Cham: Springer; 2020.
35. Uvhagen H, Hasson H, Hansson J, von Knorring M. What happened and why? A programme theory-based qualitative evaluation of a healthcare-academia partnership reform in primary care. BMC Health Serv Res 2019 Nov 01;19(1):785 [FREE Full text] [doi: 10.1186/s12913-019-4665-1] [Medline: 31675956]
36. DeLone WH, McLean ER. The DeLone and McLean model of information systems success: a ten-year update. J Manag Inf Syst 2014 Dec 23;19(4):9-30. [doi: 10.1080/07421222.2003.11045748]
37. Monitoring and Evaluating Digital Health Interventions: A Practical Guide to Conducting Research and Assessment. Geneva: World Health Organisation; 2016.
38. Goeschel CA, Weiss WM, Pronovost PJ. Using a logic model to design and evaluate quality and patient safety improvement programs. Int J Qual Health Care 2012 Aug;24(4):330-337. [doi: 10.1093/intqhc/mzs029] [Medline: 22745358]
39. Ojo AI. Validation of the DeLone and McLean information systems success model. Healthc Inform Res 2017 Jan;23(1):60-66 [FREE Full text] [doi: 10.4258/hir.2017.23.1.60] [Medline: 28261532]
40. Nigeria Health Facility Registry. Federal Ministry of Health. 2019. URL: https://hfr.health.gov.ng/facilities/hospitals-list [accessed 2021-06-21]
41. Lagos State Maternal and Child Mortality Reduction (MCMR) program: good practices memo. Lagos State Ministry of Health. URL: https://health.lagosstate.gov.ng/lagos-state-maternal-and-child-mortality-reduction-mcmr-program/ [accessed 2021-06-23]
42. Global Challenges Research Fund. UK Research and Innovation. 2021. URL: https://www.ukri.org/our-work/collaborating-internationally/global-challenges-research-fund/ [accessed 2021-03-22]
43. Purposive sampling. Lund Research. 2012. URL: http://dissertation.laerd.com/purposive-sampling.php [accessed 2017-02-20]
44. Breuer E, Lee L, De Silva M, Lund C. Using theory of change to design and evaluate public health interventions: a systematic review. Implement Sci 2016 May 06;11(1):63 [FREE Full text] [doi: 10.1186/s13012-016-0422-6] [Medline: 27153985]
45. Smith SW, Koppel R. Healthcare information technology's relativity problems: a typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ. J Am Med Inform Assoc 2014;21(1):117-131 [FREE Full text] [doi: 10.1136/amiajnl-2012-001419] [Medline: 23800960]
46. Scott PJ, Briggs JS. STAT-HI: a socio-technical assessment tool for health informatics implementations. Open Med Inform J 2010;4:214-220 [FREE Full text] [doi: 10.2174/1874431101004010214] [Medline: 21603280]
47. Fritz F, Tilahun B, Dugas M. Success criteria for electronic medical record implementations in low-resource settings: a systematic review. J Am Med Inform Assoc 2015 Mar;22(2):479-488. [doi: 10.1093/jamia/ocu038] [Medline: 25769683]
48. Cordeiro L, Soares CB. Action research in the healthcare field: a scoping review. JBI Database System Rev Implement Rep 2018 Apr;16(4):1003-1047. [doi: 10.11124/JBISRIR-2016-003200] [Medline: 29634517]
49. Thobias J, Kiwanuka A. Design and implementation of an m-health data model for improving health information access for reproductive and child health services in low resource settings using a participatory action research approach. BMC Med Inform Decis Mak 2018 Jun 25;18(1):45 [FREE Full text] [doi: 10.1186/s12911-018-0622-x] [Medline: 29941008]
50. OpenMRS homepage. OpenMRS. 2016. URL: https://openmrs.org/ [accessed 2020-03-09]
51. Muinga N, Magare S, Monda J, Kamau O, Houston S, Fraser H, et al. Implementing an open source electronic health record system in Kenyan health care facilities: case study. JMIR Med Inform 2018 Apr 18;6(2):e22 [FREE Full text] [doi: 10.2196/medinform.8403] [Medline: 29669709]

52. Purkayastha S, Allam R, Maity P, Gichoya JW. Comparison of open-source electronic health record systems based on functional and user performance criteria. Healthc Inform Res 2019 Apr;25(2):89-98 [FREE Full text] [doi: 10.4258/hir.2019.25.2.89] [Medline: 31131143]
53. Oza S, Jazayeri D, Teich JM, Ball E, Nankubuge PA, Rwebembera J, et al. Development and deployment of the OpenMRS-Ebola electronic health record system for an Ebola treatment center in Sierra Leone. J Med Internet Res 2017 Aug 21;19(8):e294 [FREE Full text] [doi: 10.2196/jmir.7881] [Medline: 28827211]
54. Welcome to FHIR®. HL7. URL: https://www.hl7.org/fhir/ [accessed 2020-10-09]
55. What is openEHR? OpenEHR. URL: https://www.openehr.org/about/what_is_openehr [accessed 2020-10-09]
56. Scott PJ, Curley PJ, Williams PB, Linehan IP, Shaha SH. Measuring the operational impact of digitized hospital records: a mixed methods study. BMC Med Inform Decis Mak 2016 Nov 10;16(1):143 [FREE Full text] [doi: 10.1186/s12911-016-0380-6] [Medline: 27829453]
57. King A, Crewe I. The Blunders of Our Governments. London: Oneworld Publications; 2014.
58. Chibanda D, Verhey R, Munetsi E, Cowan F, Lund C. Using a theory driven approach to develop and evaluate a complex mental health intervention: the friendship bench project in Zimbabwe. Int J Ment Health Syst 2016;10:16 [FREE Full text] [doi: 10.1186/s13033-016-0050-1] [Medline: 26933448]
59. CDC in Nigeria. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/globalhealth/countries/nigeria/default.htm [accessed 2022-03-22]
60. Verma N, Mamlin B, Flowers J, Acharya S, Labrique A, Cullen T. OpenMRS as a global good: impact, opportunities, challenges, and lessons learned from fifteen years of implementation. Int J Med Inform 2021 May;149:104405. [doi: 10.1016/j.ijmedinf.2021.104405] [Medline: 33639327]
61. Schade CP, Sullivan FM, de Lusignan S, Madeley J. e-Prescribing, efficiency, quality: lessons from the computerization of UK family practice. J Am Med Inform Assoc 2006;13(5):470-475 [FREE Full text] [doi: 10.1197/jamia.M2041] [Medline: 16799129]
62. Benson T. Why general practitioners use computers and hospital doctors do not--Part 1: incentives. BMJ 2002 Nov 09;325(7372):1086-1089 [FREE Full text] [doi: 10.1136/bmj.325.7372.1086] [Medline: 12424171]
63. Adler-Milstein J, Everson J, Lee SD. EHR adoption and hospital performance: time-related effects. Health Serv Res 2015 Dec;50(6):1751-1771 [FREE Full text] [doi: 10.1111/1475-6773.12406] [Medline: 26473506]
64. Nuckols TK, Smith-Spangler C, Morton SC, Asch SM, Patel VM, Anderson LJ, et al. The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. Syst Rev 2014 Jun 04;3(1):56 [FREE Full text] [doi: 10.1186/2046-4053-3-56] [Medline: 24894078]
65. Murphy EV. Clinical decision support: effectiveness in improving quality processes and clinical outcomes and factors that may influence success. Yale J Biol Med 2014 Jun;87(2):187-197 [FREE Full text] [Medline: 24910564]
66. Martin PM, Sbaffi L. Electronic health record and problem lists in Leeds, United Kingdom: variability of general practitioners' views. Health Informatics J 2020 Sep;26(3):1898-1911 [FREE Full text] [doi: 10.1177/1460458219895184] [Medline: 31875417]
67. Wallis L, Blessing P, Dalwai M, Shin SD. Integrating mHealth at point of care in low- and middle-income settings: the system perspective. Glob Health Action 2017 Jun 25;10(sup3):1327686 [FREE Full text] [doi: 10.1080/16549716.2017.1327686] [Medline: 28838302]
68. Fraser H, Adedeji T, Amendola P. The importance of primary care records in low- and middle-income settings for care, resource management and disease surveillance: a review. In: ICT in Health: Survey on the Use of Information and Communication Technologies in Brazil Healthcare Facilities. Brazil: Brazilian Network Information Center; 2021.
69. Ziemann M, Erikson C, Krips M. The use of medical scribes in primary care settings: a literature synthesis. Med Care 2021 Oct 01;59(Suppl 5):S449-S456 [FREE Full text] [doi: 10.1097/MLR.0000000000001605] [Medline: 34524242]
70. Howard K, Helé K, Salibi N, Wilcox S, Cohen M. Adapting the EHR scribe model to community health centers: the experience of Shasta community health center's pilot. Blue Shield Foundation of California. URL: https://blueshieldcafoundation.org/publications/adapting-ehr-scribe-model-to-community-health-centers-experience-shasta-community [accessed 2020-02-03]
71. Kaufman DR, Sheehan B, Stetson P, Bhatt AR, Field AI, Patel C, et al. Natural language processing-enabled and conventional data capture methods for input to electronic health records: a comparative usability study. JMIR Med Inform 2016 Oct 28;4(4):e35 [FREE Full text] [doi: 10.2196/medinform.5544] [Medline: 27793791]
72. Clinical data capture: OMR and OCR and your flatbed scanner. Medscape. URL: https://www.medscape.com/viewarticle/497865_2 [accessed 2020-08-19]
73. Bartlett L, Avery L, Ponnappan P, Chelangat J, Cheruiyot J, Matthews R, et al. Insights into the design, development and implementation of a novel digital health tool for skilled birth attendants to support quality maternity care in Kenya. Fam Med Community Health 2021 Aug 03;9(3):e000845 [FREE Full text] [doi: 10.1136/fmch-2020-000845] [Medline: 34344764]

Abbreviations

ANC: antenatal care

D&M: DeLone and McLean
EHR: electronic health record
GP: general practitioner
HCP: health care professional
HIC: high-income country
IS: information systems
LMIC: low- and middle-income country
MCH: maternal and child health care
OMR: optical mark recognition
OpenMRS: Open Medical Records System
PHC: primary health center
ToC: theory of change
WHO: World Health Organization

Edited by C Lovis; submitted 14.09.21; peer-reviewed by A Kanter, A Garcia; comments to author 31.01.22; revised version received 25.03.22; accepted 31.05.22; published 11.08.22.

Please cite as:
Adedeji T, Fraser H, Scott P
Implementing Electronic Health Records in Primary Care Using the Theory of Change: Nigerian Case Study
JMIR Med Inform 2022;10(8):e33491
URL: https://medinform.jmir.org/2022/8/e33491
doi:10.2196/33491
PMID:35969461

Original Paper

Electronic Data Capture System (REDCap) for Health Care Research and Training in a Resource-Constrained Environment: Technology Adoption Case Study

Irma Adele Maré[1,2], BScHons, MSc, MSc (Med); Beverley Kramer[3], BSc (Hons), PhD; Scott Hazelhurst[2,4,5], BScHons, MSc, PhD; Mapule Dorcus Nhlapho[2,4,6], BTech; Roy Zent[7], MD, PhD; Paul A Harris[1,8,9,10], PhD; Michael Klipin[1,2], BSc (Hons), MBBCh

[1]Department of Surgery, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[2]Division of Biomedical Informatics and Translational Science, Wits Health Consortium, Johannesburg, South Africa

[3]School of Anatomical Sciences, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[4]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[5]School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

[6]Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[7]Division of Nephrology and Hypertension, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

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

[9]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, United States

[10]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Irma Adele Maré, BScHons, MSc, MSc (Med)
Department of Surgery, School of Clinical Medicine
Faculty of Health Sciences
University of the Witwatersrand
1 Jan Smuts Avenue
Braamfontein
Johannesburg, 2000
South Africa
Phone: 27 0117171000
Email: aicm.v2@gmail.com

Abstract

Background: Electronic data capture (EDC) in academic health care organizations provides an opportunity for the management, aggregation, and secondary use of research and clinical data. It is especially important in resource-constrained environments such as the South African public health care sector, where paper records are still the main form of clinical record keeping.

Objective: The aim of this study was to describe the strategies followed by the University of the Witwatersrand Faculty of Health Sciences (Wits FHS) during the period from 2013 to 2021 to overcome resistance to, and encourage the adoption of, the REDCap (Research Electronic Data Capture; Vanderbilt University) system by academic and clinical staff. REDCap has found wide use in varying domains, including clinical studies and research projects as well as administrative, financial, and human resource applications. Given REDCap's global footprint in >5000 institutions worldwide and potential for future growth, the strategies followed by the Wits FHS to support users and encourage adoption may be of importance to others using the system, particularly in resource-constrained settings.

Methods: The strategies to support users and encourage adoption included top-down organizational support; secure and reliable application, hosting infrastructure, and systems administration; an enabling and accessible REDCap support team; regular hands-on training workshops covering REDCap project setup and data collection instrument design techniques; annual local symposia to promote networking and awareness of all the latest software features and best practices for using them; participation in REDCap Consortium activities; and regular and ongoing mentorship from members of the Vanderbilt University Medical Center.

Results: During the period from 2013 to 2021, the use of the REDCap EDC system by individuals at the Wits FHS increased, respectively, from 129 active user accounts to 3447 active user accounts. The number of REDCap projects increased from 149 in 2013 to 12,865 in 2021. REDCap at Wits also supported various publications and research outputs, including journal articles and postgraduate monographs. As of 2020, a total of 233 journal articles and 87 postgraduate monographs acknowledged the use of the Wits REDCap system.

Conclusions: By providing reliable infrastructure and accessible support resources, we were able to successfully implement and grow the REDCap EDC system at the Wits FHS and its associated academic medical centers. We believe that the increase in the use of REDCap was driven by offering a dependable, secure service with a strong end-user training and support model. This model may be applied by other academic and health care organizations in resource-constrained environments planning to implement EDC technology.

(JMIR Med Inform 2022;10(8):e33402) doi:10.2196/33402

KEYWORDS

electronic data capture; implementation science; Research Electronic Data Capture; REDCap; biomedical informatics; South Africa

Introduction

Background

Challenges to Electronic Data Capture Implementation in Health Care

Electronic data capture (EDC) and management is a vital part of the administrative process in almost all industries, but despite its many advantages, adoption in the health care research and clinical service delivery domains has lagged behind other industries [1-13]. Major obstacles include cost, lack of policy at management level, implementation failure, and data security concerns [1-7]. In addition, research information applications developed specifically for one environment usually require significant resources and support to function in another environment because the technology, support, security, and privacy needs are often different [5,8-10]. Many institutions in sub-Saharan Africa and other low- and middle-income countries lack the technical resources to support information systems for health care research and clinical service delivery [11-13]. Moreover, power and network infrastructure may be unreliable [11-13].

Even with the necessary organizational and infrastructure support, the success of EDC software applications is not guaranteed [4,5,14]. The lack of domain knowledge in resource-constrained environments such as sub-Saharan Africa has hampered the implementation of EDC technologies [11]. There is often a paucity of individuals with the necessary clinical, academic, and IT skills required to support critical health care data management systems [11]. Field-workers and clinician scientists, although highly skilled and valued in their respective domains, may not be well versed in technology for the capture, storage, and transmission of health data [4-7].

Once infrastructure and skills-resourcing issues have been overcome, familiarity with deeply ingrained systems and processes at every level of the research enterprise is a natural cause of resistance to change [1,4-7,15-18]. This includes, for example, the replacement of hard-copy files with electronic data collection instruments for clinical research informatics. Strategies to obviate the resistance to implementation of new technology include demonstrating trustworthiness and benefits of the technology, while at the same time easing the transition through access to appropriate training and support [1,5-7,16,17].

The State of EDC at the University of the Witwatersrand Faculty of Health Sciences Before Research Electronic Data Capture Adoption

The University of the Witwatersrand (Wits) is a research-intensive university based in the metropolitan area of Johannesburg, South Africa, an upper middle–income environment [19]. The Wits Faculty of Health Sciences (FHS) operates within a health care system weakened by sociopolitical and historical issues and strained by an ongoing quadruple burden of disease [12,20,21]. Setting up systems to support academics in their clinical and research activities is subject to budget limitations and constrained by the geographical distribution of the approximately 2500 health sciences staff and 7000 students, spread over 3 discrete academic teaching platforms and many field sites in both urban and rural regions [22,23].

Before the implementation of the centrally supported EDC system at the University of the Witwatersrand Faculty of Health Sciences (Wits FHS) and the associated research entities in 2012, electronic data management services were fragmentated, inconsistent, and variable. Individuals and research entities were using local devices and legacy systems familiar to them or choosing new products to implement based only on their own needs, abilities, and budget. This fragmentation was not desirable from an organizational perspective because the data sets were isolated, the financial and human resources used for procurement and management were diluted, and the security and privacy of data could not be guaranteed [5,9,10].

During the same period, most of the patient health record data being collected at Wits FHS–related medical centers were paper-based. Clinical staff were burdened by service delivery demands [20] and restricted in the time they had available for research or data capture [24]. Where electronic research data collection instruments were used, they ranged from simple spreadsheet programs such as Microsoft Excel to enterprise-wide data management systems. At the time, there were no preferred instruments within the Wits FHS, and the safety, security, and privacy of the data were at the discretion of individual

researchers. There was little standardization of metadata or clinical coding systems, and, as a result, interoperability and secondary analysis of data were rare. This affected patient care and limited the dissemination of important knowledge gained in treating various infectious diseases and diseases of lifestyle that continue to burden the South African health care system [12,20,21].

Prior Work: Implementation of Research Electronic Data Capture at Wits FHS (2012-2013)

In 2012, as part of a process to strengthen research support, training, and outputs [8,24], the Wits FHS implemented REDCap (Research Electronic Data Capture; Vanderbilt University), a web-based EDC tool created by informaticists at the Vanderbilt University Medical Center (VUMC) in Nashville, Tennessee [8,25].

REDCap allows users to build electronic data collection instruments for a wide range of data types and environments. It is specifically geared toward research studies and operational data and toward enabling the capture and management of data in a manner that is compliant with 21 Code of Federal Regulations Part 11, the Federal Information Security Management Act, the Health Insurance Portability and Accountability Act, and General Data Protection Regulation [14,25]. Each project will have its own procedures for validation and quality control, and REDCap has many features available to end users to support this, such as granular user rights, a detailed audit log, and data quality control tools

One of the strategic goals of the Wits FHS was to unify and systematize health care and research data collection within the institution [8]. REDCap was an attractive option because of its freeware licensing model for noncommercial use and large international support community [14,25]. The decision to implement REDCap was also supported by an existing diaspora relationship between the Wits FHS and the VUMC [26].

The strategies used to install REDCap and overcome the initial implementation barriers within the Wits FHS have been presented previously [8]. The crucial factors highlighted in the paper were support from the Wits FHS management and the allocation of a modest budget for hosting infrastructure, systems administration, and recruitment of personnel from existing staff for end-user support. Support staff were initially allocated part time on a sliding scale of need, which allowed dedicated end-user support while limiting costs. The hardware costs were limited to servers and security certificates, which were housed at existing university data centers. Four months after the implementation, the number of REDCap users at the Wits FHS was 81, and after 12 months it had increased to 140. The total costs to provide a functional REDCap platform for the first year was <US $9000 [8].

Goal of This Study

EDC implementation projects often fail after the initial deployment because of resistance from, and lack of adoption by, end users, even when leadership, infrastructure, and human resources are mobilized successfully [1,2,5,9,13,18]. The Wits FHS used various strategies to engage and support end users to overcome these challenges, and the period from 2013 to 2021 was characterized by sustained—sometimes exponential—growth in the demand for REDCap accounts and support services. The aim of this paper was to discuss the methods that helped to overcome barriers to adoption because we believe that these strategies may be applied in other low- to middle-income and resource-constrained environments where EDC implementation and adoption are subject to similar challenges. We measured the growth in REDCap use by increases in user accounts, projects, and publication metrics at the Wits FHS. The key success factors identified and discussed in detail in this paper are as follows:

- Top-down organizational support for EDC
- A proactive response and support team that can train and support users
- Continual development of the support team through mentoring and participation in national and international activities
- Maintaining visibility through promotion campaigns, networking events, and academic symposia
- Collaborating with, and learning from, established international partners
- Secure and reliable hosting

Methods

Adoption-Support Strategies

Hosting and Systems Administration

To gain the trust of users, the reliability of the Wits REDCap system was paramount. The system was deployed on 2 virtual machines—one for the MySQL database and a second one for the REDCap application—on a reliable Intel server with 64 GB of RAM and considerable disk space. The Ubuntu long-term support operating system version current at the time was used for both virtual host and physical machines. New hardware was introduced every 3 to 4 years; the cost implications when amortized over the life span of the machine were small. Older, retired hardware was recycled for less-critical work. Multiple levels of backup were used. Daily backups of the database and uploaded files were kept on a separate machine. Three times a week copies of the virtual machine images were created and stored on a server on a different campus. No major hardware failures occurred during this period, but tests were performed to emulate recovery using the backed-up data to ensure that this would be possible in the event that the primary hardware failed. Nagios software (Nagios Enterprises, LLC) [27] was used to monitor system health and stability. The electricity supply in South Africa was periodically unreliable; however, the server was placed in the university's data center, with multiple backup power redundancies and physical security infrastructure. The REDCap application itself proved to be highly reliable, with regular bug fixes, security, and functionality updates released by the VUMC developers. The human resource allocation dedicated to system administration from 2013 to 2019 was equal to approximately 0.05 full-time equivalents (FTEs).

In 2019, the load created by concurrent users and processes made it necessary to move from magnetic to solid state drives. From 2020 onward, infrastructure and systems administration

demands outgrew the existing resources. The Wits FHS REDCap system was moved from the Wits data center to a leading South African cloud hosting provider (Teraco) [28]. Additional systems administration capacity from the Wits Health Consortium [29], a wholly owned subsidiary of Wits, was brought in to manage the cloud hosting environment, with continued guidance and leadership from the original systems administrator. The Wits Health Consortium infrastructure team uses Arcserve [30] for daily snapshots of the virtual environment, with a separate backup every 1 to 3 days on a removable storage device for offsite storage. At the time of writing, the time spent on systems administration totaled approximately 0.1 FTE.

End-User Support

The second crucial component of the Wits FHS implementation strategy was a dedicated *go-to* individual to support end users, known as the REDCap administrator [8]. One of the most effective types of individuals to place in this role is referred to in the literature as a "technology bridger" [2,9,31]. A bridger is an early adopter who has a deep understanding of the technology being implemented as well as the soft skills to teach and support others at their organization.

The implementation of technology is a form of change management, and by approaching end-user support with an open-door policy and a culture of *psychological safety* ("a belief that one will not be punished or humiliated for speaking up with ideas, questions, concerns, or mistakes" [32]), the anxiety concerning, and resistance to, change exhibited by end users is reduced [personal communication by Wits REDCap administrator, July 2021].

By encouraging one-on-one consultations with the REDCap administrator in a relaxed and informal setting, new users felt safe to discuss their concerns or expose where they might have a lack of understanding. All email support and one-on-one consultations were provided at no cost. Over time, the need for additional project design and management services for larger and more complex projects became clear. To protect the REDCap administrator from users wanting to make use of the design service, rather than engaging with the administrator, to learn how to use the system on their own, an hourly fee was implemented for design services.

As the number of end users grew, so did the support and administrative needs (Textbox 1). During the period 2013-2014, a part-time REDCap administrator (0.5 FTE) was adequate. This was increased to 1.0 FTE from 2014, and a second full-time administrator was added in 2016. Additional part-time REDCap administrators were added in 2021 in response to a large increase in system use, currently totaling approximately 2.2 FTEs on average.

Textbox 1. Number of REDCap (Research Electronic Data Capture) application administrator full-time equivalents (FTEs) per year.

REDCap application administrator FTEs

- 2013: 0.5
- 2014: 1.0
- 2015: 1.0
- 2016: 2.0
- 2017: 2.0
- 2018: 2.0
- 2019: 2.0
- 2020: 2.0
- 2021: 2.2

Hands-on REDCap Training

The REDCap application has a number of built-in tutorial videos and extensive *Help* and *Frequently Asked Questions* documentation. However, our experience showed that in addition to one-on-one consultations, the majority of late-adopting end users benefited from in-person formal participative instruction in the use of REDCap design tools. A series of sessions were offered in 2013, the content and format of which informed the creation of formal structured REDCap training workshops in 2014 and thereafter. Although the workshops were well attended, there was a significant proportion of attendees who had made reservations but did not attend. This prompted the introduction of a registration fee, which improved compliance and provided funds to contribute to the sustainability of the REDCap support team.

Initially, a basic introduction-to-REDCap workshop was offered, but as users became more skilled, a more advanced workshop was added in 2015. Good design practices and standardization of metadata were encouraged, and the workshops also established an interpersonal relationship between the end users and the REDCap administration team. A few groups, both within and external to the Wits FHS, requested on-demand REDCap training programs similar in content to that of the workshops. The number of REDCap training workshops and attendees are summarized in Table 1 and Table 2, respectively.

Each year, approximately 5 introductory and 4 advanced workshops were delivered, with an average of 19 and 8 attendees at each type of workshop, respectively. Step-by-step workshop manuals were also compiled iteratively over time and provided to attendees from 2017 onward. A total of 977 individuals attended all workshops between 2014 and 2020: 721 (73.8%)

were Wits FHS–affiliated, whereas 256 (26.2%) were from external organizations. Fewer workshops were offered in 2018-2019 because of staffing constraints. From March 2020 onward, workshops were shifted to an internet-based platform (Zoom; Zoom Video Communications, Inc) because of COVID-19 restrictions on in-person gatherings. The availability of web-based teaching resulted in an increase in on-demand workshops—5 were held in 2020 compared with between 1 and 3 in previous years. We attribute this increase in part to the wider access inherent in the web-based format and partly because of the increased number of new users in 2020.

Table 1. The number of REDCap (Research Electronic Data Capture) workshops from 2014 to 2020, as recorded by web-based booking forms and attendance registers.

Workshops	2014	2015	2016	2017	2018	2019	2020
Introductory	2	6	6	7	5	5	5
Advanced[a]	0	2	3	5	3	4	3
On-demand	0	2	1	1	3	0	5
Total	2	10	10	13	11	9	13

[a]Before 2019, these workshops were referred to as *intermediate* hands-on *REDCap workshops*. A significant proportion of novice users attempted the intermediate sessions and found the content and pace of the intermediate session to be beyond their capacity. We therefore changed the name to *advanced* hands-on *REDCap workshops* and set up entry requirements in 2019 to emphasize that attendees had to have mastered the basics on their own or attended an introductory session before attempting the advanced one.

Table 2. The number of REDCap (Research Electronic Data Capture) attendees from 2014 to 2020, as recorded by web-based booking forms and attendance registers.

Attendees	2014	2015	2016	2017	2018	2019	2020
Wits FHS[a]	21	108	167	110	113	94	108
External	9	38	45	80	25	16	43
Total	30	146	212	190	138	110	151

[a]Wits FHS: University of the Witwatersrand Faculty of Health Sciences.

REDCap Consortium Participation

Overview

The *REDCap Consortium* is a community of REDCap administrative and technical support staff from the academic, nonprofit, and government institutions that have adopted REDCap [14]. The REDCap Consortium represents a *professional home* for many of the local REDCap research informatics leaders, and it is a forum for enabling teams to share ideas, problems, and solutions related to the innovative use of REDCap [14]. Feedback and communication with the local administrators participating in the REDCap Consortium is a vital component of how REDCap program leaders understand unmet needs, socialize concepts for new features, and eventually prioritize new development. The voluntary participation in the REDCap Consortium is a very valuable investment for organizations [14,33]. Membership of the consortium gives partner organizations' REDCap administrators access to various networking and information-sharing platforms, of which there are three main types: (1) consortium calls, (2) conferences and symposia, and (3) the REDCap Community forum. Between late 2013 and April 2020, the number of REDCap Consortium members from Africa and South Africa grew from 26 and 8, respectively, to 261 and 175, respectively [8,34].

REDCap Consortium Calls

The REDCap software development team at the VUMC hosts a weekly technical call, a forum to share news and updates and generally bring the REDCap developer and administrator community up to date with the latest REDCap features. Since 2016, the VUMC has also hosted 2 Eastern Hemisphere Partner consortium calls at times that made them accessible to the Africa and Europe as well as Australia, New Zealand, and Japan regions. Various subcommittees have also been formed within the REDCap Consortium, either based on common interest (software validation or development of training materials) or on shared geographical location and language (Hispanophone or Francophone committees). Some of these subcommittees also host regular calls relating to their specific domains. A locally hosted call for the African region was added in 2018 through a collaboration between the REDCap administrator of the Pan-African Bioinformatics Network for Human Heredity and Health in Africa, based at the University of Cape Town, and the Wits REDCap team. Participation in these calls supports and develops REDCap administrators by keeping them up to date with the latest developments, enabling networking and exchange of ideas, as well as giving them a platform to connect directly with the REDCap software development team at the VUMC.

Conferences and Symposia

The annual REDCap conference (REDCapCon) is a forum for REDCap administrators from different countries, institutions, and environments to meet, share experiences, and create a support network. The opportunity to interact with international members forms the basis of a collective resource for information dissemination and problem solving within the global REDCap Community. Participation in the annual REDCapCon has proved

very valuable in terms of the opportunity to attend and present our work. A representative from Wits FHS has attended the annual REDCapCon since 2015. However, many sub-Saharan Africa–based REDCap administrators do not have a budget to travel to North America for the annual REDCapCon. It became apparent that an African REDCapCon would add value to the African consortium partners. Wits hosted the first REDCap Africa Day in Johannesburg in 2016 as an adjunct to the FHS research day, followed by 3 more REDCap Africa symposia in 2017, 2019, and 2020 (Table 3). Each symposium has been attended by one or more members of the VUMC REDCap team. The REDCap Africa event encourages attendance by regional and international REDCap administrators as well as end users, as opposed to the REDCapCon, where only administrators attend. The REDCap Africa symposia agendas were a mixture of technical presentations and use cases, with plenary sessions presented by local academics as well as VUMC visitors. Delegates included local and regional faculty members, research institute employees, and students.

The novelty of the first REDCap Africa symposium drew large numbers of attendees (Table 3), including a number of casual attendees from the Wits FHS who were not REDCap users but nonetheless wanted to learn more about REDCap, the EDC strategy of the Wits FHS, and the relationship with the VUMC. In subsequent years, the number of casual attendees decreased and was made up of REDCap administrators or highly engaged power users. The cost of intra-Africa travel is high, and the number of attendees from the African region who were able to attend in person remained low. In 2020, because of COVID-19 restrictions, REDCap Africa Day was hosted using Zoom, and attendance was significantly higher with more international delegates than at any previous event. In October 2021, again because of the ongoing COVID-19 pandemic, we organized another Zoom-based installment of REDCap Africa Day (we have institutional review board clearance to report data up to September 30, 2021; hence, we cannot report the actual number of attendees), but future REDCap Africa symposia will explore combined in-person and livestreaming as well as travel bursaries to encourage regional participation.

Table 3. The number of delegates to the REDCap (Research Electronic Data Capture) Africa symposia per year.

	2016	2017	2019	2020
Wits FHS[a] affiliated delegates	69	40	27	42
Local (SA[b]) delegates	30	28	19	37
International delegates	6	2	4	31
Total number of delegates	105	70	50	110

[a]Wits FHS: University of the Witwatersrand Faculty of Health Sciences.

[b]SA: South Africa.

REDCap Community Forum

The REDCap Community is a web-based platform where the administrative and IT support staff of a consortium partner institution can access software downloads, extensive technical documentation, a question-and-answer forum, consortium announcements, committee activities, events, and more [35]. The REDCap Community website provides a forum for interaction on, and dialogue about, REDCap-related topics with REDCap administrators around the globe, and it is an essential resource for the development of an institution's capacity to host and support REDCap. The Wits FHS REDCap administrators have received technical advice or otherwise benefited from discussions on the REDCap Community forum, while also enjoying the networking experience and the sense of community gained by interacting with peers from other institutions.

VUMC Relationship and Support

The development of clinical research informatics capacity and health care IT skills are particularly important in resource-constrained settings such as sub-Saharan Africa, and the Wits-VUMC partnership has contributed to the capacitation process at the Wits FHS. Initially, the relationship between Wits and the VUMC grew from an alumni diaspora program initiated in 2010 [36]; it later expanded through bilateral visits by academic staff [26], custom REDCap development projects, joint grant applications, and mentorship. New REDCap Consortium members in resource-constrained environments may benefit from leveraging existing diaspora linkages or from initiating new mentorship and capacity-development collaborations with institutions, such as the VUMC, that have mature research informatics divisions.

Measurement of System Use and Growth

Overview

REDCap system use is reported in 2 main ways: the number of users and the number of projects. The number of users and projects on any REDCap system is available on the application's administration web page called the *Control Center*. The Wits FHS REDCap administrator has a monthly record of several use metrics since system installation in August 2012. Other metrics were determined retrospectively by running MySQL queries on the REDCap database logs with the help of a REDCap external module named *MySQL Simple Admin* [37]. These queries were used to identify and enumerate project- or user account–creation events. The descriptive system-use metrics that we report on are as follows:

1. Total number of user accounts
2. Number of user accounts that were active each year
3. Annual increase or decline in number of active user accounts
4. Total number of practice and nonpractice projects
5. Project-purpose attribute of each nonpractice project

User Accounts

A REDCap user account allows an individual user to access their own private repository of projects and data. The REDCap Control Center displays various metrics regarding accounts; for example, the total number of user accounts, the number of accounts that were active in a given period, and the number of accounts that were suspended because of inactivity. Activity is defined as logging in and performing an action on REDCap. At the Wits FHS, the period of inactivity that leads to suspension is 180 days (this may differ from institution to institution based on policy). We used a combination of Control Center statistics and MySQL Simple Admin queries to report the total number of user accounts on the Wits FHS REDCap system per year as well as the number of accounts that were active in a given year. The measurement is taken on the first day of September each year.

In addition, the growth and decline in the number of active user accounts were determined by taking the number of active users for a given year and subtracting the number of active users from the previous year.

Projects

A *project* in REDCap refers to a set of connected electronic data collection screens and records that are related to a specific purpose.

When creating a project, end users are required to allocate one of five possible project purposes, namely *practice*, *research*, *quality improvement*, *operational support*, or *other*. The Control Center report usually excludes practice projects from the *total projects* metric, but we performed a query using MySQL Simple Admin to retrieve the number of practice projects as well, and we report these together with the total nonpractice projects. We believe that practice projects are an important metric of how comfortable users are to experiment and explore the system, something which is actively encouraged by our REDCap administrators during one-on-one or group training sessions.

Furthermore, we performed an annual breakdown of the nonpractice projects by purpose, and report the percentage of projects in the categories of *research*, *quality improvement*, *operational support* and *other*.

Measurement of System-Related Research Outputs

One way to measure the impact of REDCap on the Wits FHS is by reviewing research outputs such as journal articles and postgraduate theses. We performed a bibliometric survey to determine the number of research outputs from Wits FHS–affiliated authors that relied on the use of Wits FHS REDCap for EDC or data management and report the number of outputs per year as well as the subject domains across all the outputs for the period from 2013 to 2020. REDCap has a built-in publication-matching tool that will review PubMed databases for authors and affiliations that match those of REDCap users or project principal investigators. We evaluated the list of *potential matches* generated within REDCap and only included articles and monographs in our results if they mentioned using Wits FHS REDCap as the data collection instrument in the methods or acknowledgments sections. In cases where no system was mentioned by name, we contacted the authors through email to request clarification and only included the output if authors confirmed in writing that Wits FHS REDCap was used.

Ethics Approval

Permission to perform the research was obtained from the Wits Department of Surgery postgraduate protocol committee, the Wits Human Research Ethics Committee (M210551), and the Wits University Registrar.

Results

User Accounts

There was a sustained—and at times exponential—increase in the number of user accounts from 139 total and 129 active accounts in 2013 to 7128 total and 3447 active accounts in 2021 (Figure 1).

The number of active users has increased 25-fold in 8 years; however, the magnitude of the growth was variable (Figure 2). Except for the 2016 and 2018-2019 periods, annual growth exceeded 25%.

Figure 1. User account statistics for the University of the Witwatersrand Faculty of Health Sciences REDCap (Research Electronic Data Capture) system from September 2013 to September 2021. The total number of accounts as well as the active cohort is shown for each year.

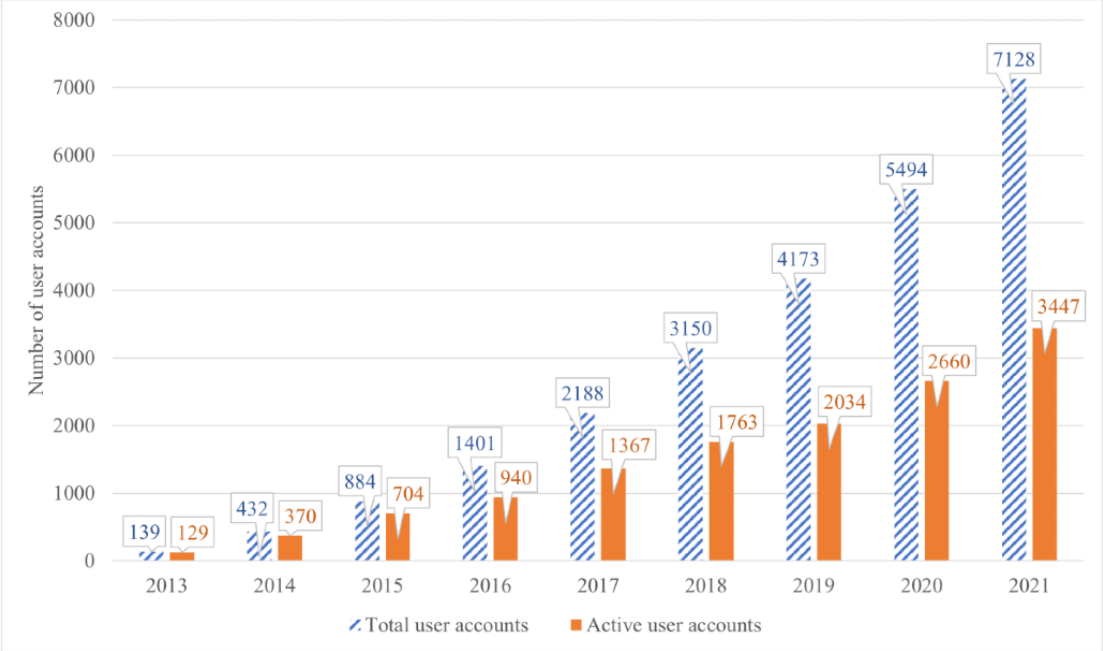
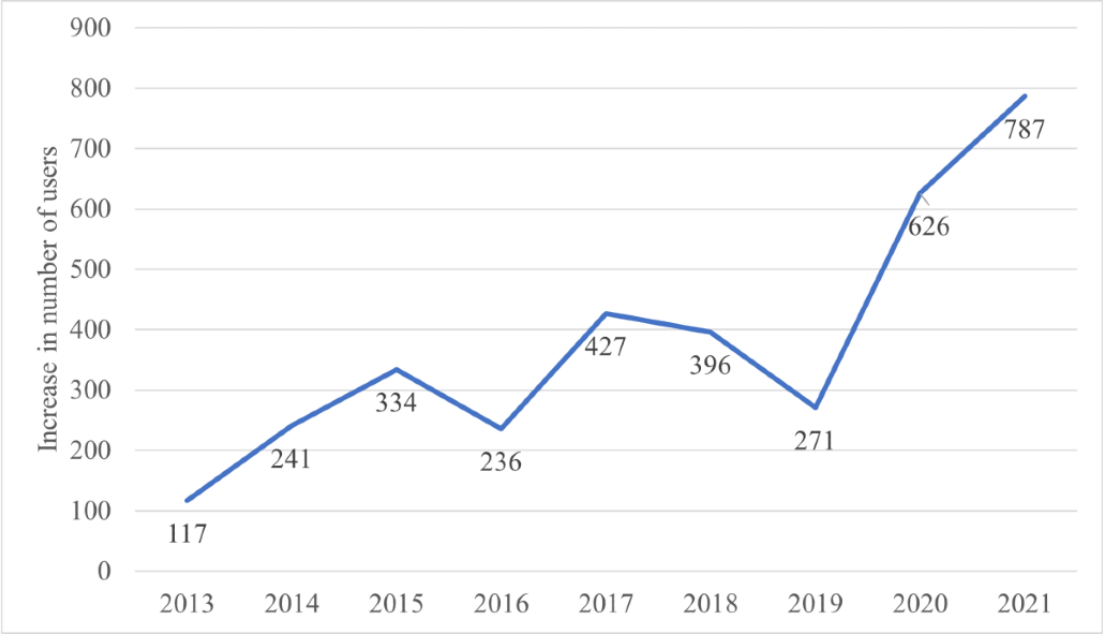


Figure 2. The increase in the number of active REDCap (Research Electronic Data Capture) users at the University of the Witwatersrand Faculty of Health Sciences from 2013 to 2021. The values were obtained by taking the number of active users in a given year and subtracting the number of active users recorded for the previous year.



Projects

The total number of projects on the Wits REDCap platform increased significantly from 149 in September 2013 to 12,865 in September 2021 (Figure 3). The number of nonpractice projects increased from 97 in September 2013 to 7038 in September 2021 and accounted for 54.71% (7038/12,865) of the total number of projects in 2021.

Of the 7038 nonpractice projects on the Wits REDCap system in September 2021, the majority (n=3952, 56.15%) were for research purposes (Table 4), and the remaining (n=3086, 43.85%) were dedicated to nonresearch purposes (Table 4).

The majority of the Wits REDCap nonresearch projects in 2021 were allocated as *operational support* (1850/3086, 59.95%), whereas *quality improvement* and *other* represented a minority (705/3086, 22.84%, and 531/3086, 17.21%, respectively).

From Table 4, it can be seen that the *operational support* category has grown to represent a larger share of the total projects every year (from 10/97, 10%, in 2013 to 1850/7038, 26.29%, in 2021), whereas the research, quality improvement, and other categories have all declined as a percentage of the total.

Figure 3. REDCap (Research Electronic Data Capture) projects on the University of the Witwatersrand Faculty of Health Sciences system per annum from September 2013 to September 2021.

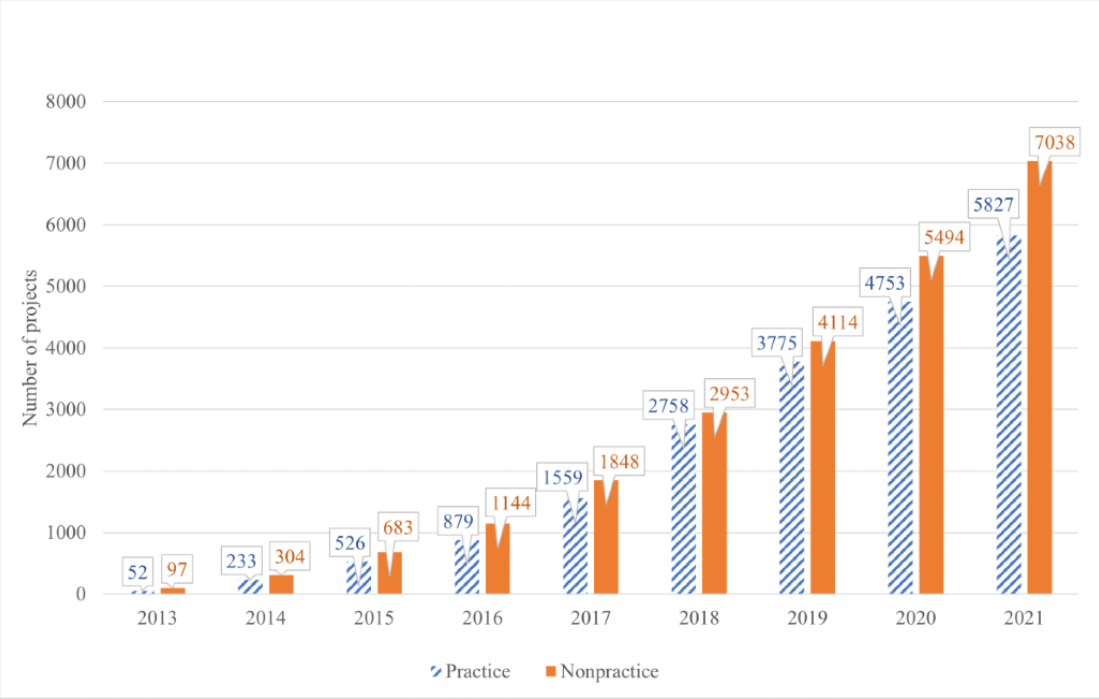


Table 4. The number of nonpractice projects on the University of the Witwatersrand Faculty of Health Sciences REDCap (Research Electronic Data Capture) system from September 2013 to September 2021, categorized by purpose.

	Research, n (%)	Operational support, n (%)	Quality improvement, n (%)	Other, n (%)
2013 (n=97)	62 (63.92)	10 (10.31)	16 (16.5)	9 (9.28)
2014 (n=304)	209 (68.75)	36 (11.84)	38 (12.5)	21 (6.91)
2015 (n=683)	452 (66.18)	79 (11.57)	88 (12.88)	64 (9.37)
2016 (n=1144)	757 (66.17)	170 (14.86)	116 (10.14)	101 (8.83)
2017 (n=1848)	1223 (66.18)	324 (17.53)	169 (9.15)	132 (7.14)
2018 (n=2953)	1842 (62.38)	593 (20.08)	265 (8.97)	253 (8.57)
2019 (n=4114)	2523 (61.33)	887 (21.56)	392 (9.53)	312 (7.58)
2020 (n=5494)	3145 (57.24)	1376 (25.05)	558 (10.16)	415 (7.55)
2021 (n=7038)	3952 (56.15)	1850 (26.29)	705 (10.02)	531 (7.54)

Wits REDCap Publication Metrics

In total, 233 journal articles and 87 postgraduate research monographs acknowledging the use of the Wits FHS REDCap system were published between 2013 and 2020. As shown in Figure 4, the number of articles increased over time as more users adopted the system and as projects reached maturity and results were disseminated. The year 2020 saw a sharp decrease in postgraduate monographs that cite the Wits FHS REDCap system. This may be due to a delay before a thesis or dissertation becomes available on the institutional repository. Additional delays in postgraduate submissions could be a result of strict COVID-19 shutdowns in South Africa.

A visualization of the research areas of the journal articles also illustrates the diversity of scientific disciplines on which REDCap at the Wits FHS had an impact (Figure 5).

Figure 4. The number of research outputs linked to use of the University of the Witwatersrand Faculty of Health Sciences REDCap (Research Electronic Data Capture) system, grouped by publication year. Postgraduate monographs include PhD and master's research works. Certain data included herein are derived from Clarivate Web of Science (copyright Clarivate 2021; all rights reserved).

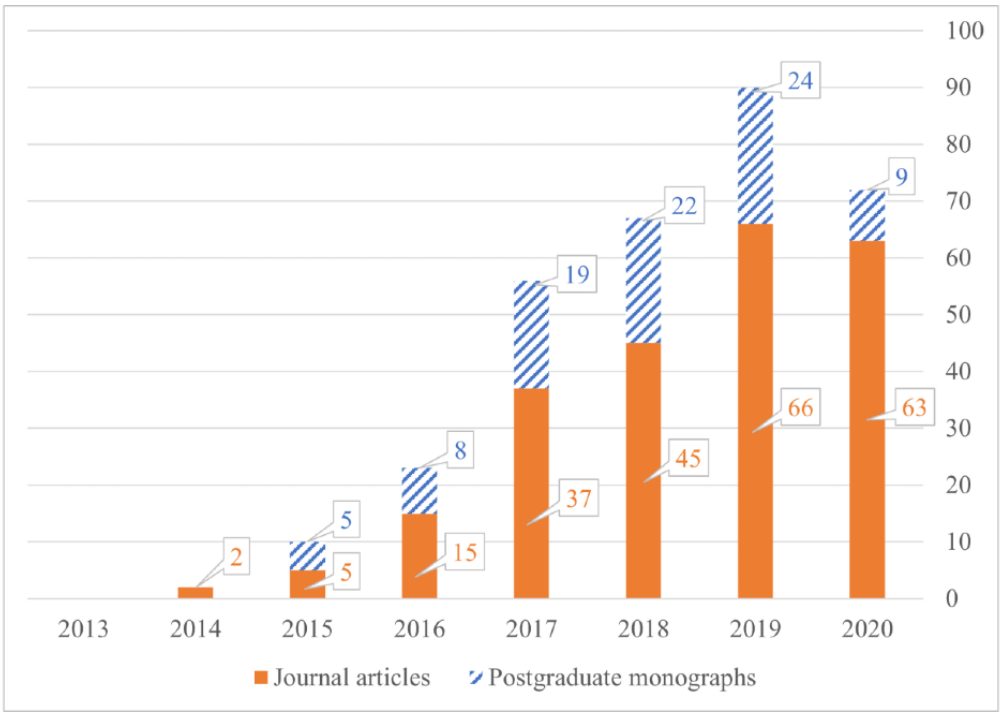
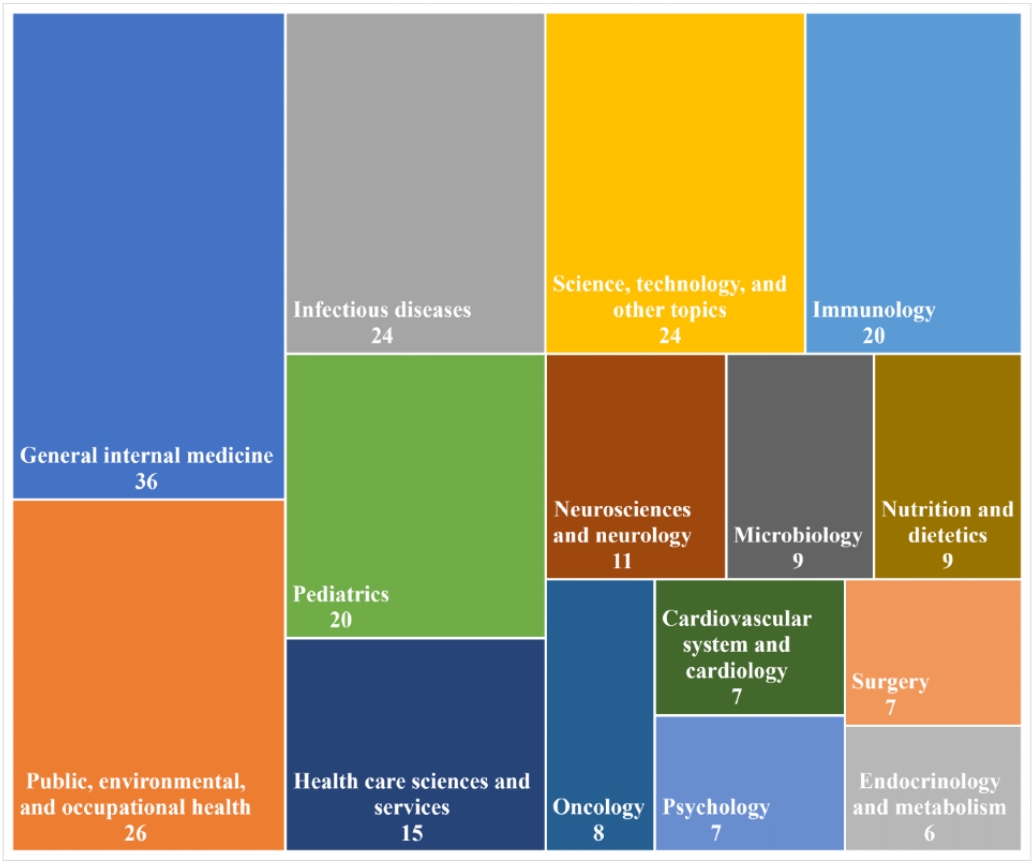


Figure 5. Treemap visualization showing the Witwatersrand Faculty of Health Sciences publications between 2013 and 2020 that were supported by REDCap (Research Electronic Data Capture), categorized by research area. This graph represents the top 15 out of 50 research areas by frequency of occurrence. Certain data included herein are derived from Clarivate Web of Science (copyright Clarivate 2021; all rights reserved).



Discussion

Principal Findings

During the period 2012-2013, the Wits FHS implemented the REDCap EDC system to support both research and clinical service delivery data management needs [8]. REDCap provides researchers with the means to design and develop EDC tools that conform with international best practices for the safety, security, and privacy of clinical data. REDCap is licensed by Vanderbilt University at no charge to government, academic, and nonprofit organizations for use in noncommercial academic and research contexts [14,25]. By removing financial burden and providing tools and support channels that empower local research informatics leaders to serve the local research enterprise, the REDCap platform fills a critical gap in most research organizations [14,25].

During the period from 2013 to 2021, use of the REDCap EDC system increased steadily at the Wits FHS, as evidenced by the growth of users, projects, and publications.

The number of active Wits FHS REDCap users increased from 129 in 2013 to 3447 in 2021, which is a 25-fold increase. Each year saw more active users than the preceding year, and with the exception of 2016, 2018, and 2019, the annual increase in the number of users was larger every year. Although our records do not contain an explanation for the slower growth in 2016, the 2018-2019 period saw fewer hands-on workshops being offered because of the senior REDCap administrator being on extended leave. In 2020 and 2021, the annual increase was 1.5-fold and 2-fold higher, respectively, than in any previous year. Two factors may have contributed to the 2020-2021 surge in active users: first, the COVID-19 pandemic drove the adoption of web-based instruments that could be accessed by teams working remotely, and second, the Protection of Personal Information Act (POPIA) was introduced in South Africa in 2020 and enforced after July 2021 [38]. As part of POPIA requirements, organizations that process personal information were required to use secure, auditable applications, and REDCap was one of only a few products offered by our institution that were compatible with POPIA requirements.

The total Wits FHS REDCap projects numbered 149 in 2013 and 12,865 in 2021. Of the 12,865 projects in 2021, a total of 5827 (45.29%) projects were created for *practice* purposes, and 7038 (54.71%) projects were for nonpractice purposes. The ratio of practice to nonpractice projects has remained remarkably stable at approximately 45% over time. Practice projects are created during the formal hands-on training workshops offered by the Wits FHS, and the use of practice projects to test design ideas and prototype data collection instruments are encouraged by the Wits FHS REDCap administrators during one-on-one consultations. The REDCap application developers release updated features on an almost monthly basis, meaning that even experienced users might resort to creating practice projects from time to time to test out new functionality. A deeper analysis of user behavior may be needed to determine conclusively the reason for the observed stability of the ratio of practice to nonpractice projects.

Of the 7038 nonpractice projects on the Wits REDCap system in September 2021, the majority were for research purposes (n=3952, 56.15%; Table 4), whereas the remaining (n=3086, 43.85%) projects were dedicated to nonresearch purposes (Table 4). This reflects the diversity of the uses of REDCap in the Wits environment: as a clinical health record and a staff rostering and management tool, as well as in a multitude of *spreadsheet*-type administrative processes, in addition to research. Further innovative off-label uses of REDCap appear in the literature [39-42].

During the period from 2013 to 2020, a total of 233 papers and 87 postgraduate monographs that acknowledge the use of Wits FHS REDCap were published. The increase in research outputs occurred during a time when the Wits FHS research and postgraduate support office used several strategies to increase research and publication rates [24], and REDCap was one of the contributing factors to the observed rise in publication metrics of Wits FHS staff and students.

EDC-Support Strategies in a Resource-Constrained Environment

The growth in the use of REDCap at Wits FHS was driven in large part by the trust generated by offering a reliable, secure service and a strong end-user training and support model (Textbox 2). A critical success factor is that hosting and server infrastructure was supported by a highly experienced systems administrator who ensured that appropriate security measures and disaster recovery plans were in place. Additional storage and computing capacity was added as needed. It was important to respond to performance problems when incremental upgrades were no longer sufficient. End users were supported by a dedicated REDCap administrator available through an email helpdesk, one-on-one consultations, formal training workshops, and annual symposia. The support team size was expanded over time to meet the increase in demand (REDCap administrator or administrators: from 0.5 FTEs in 2013 to 2.2 FTEs in 2021 and systems administrator or administrators: from 0.05 FTEs in 2013 to 0.1 FTEs in 2021). The capacity of the support staff was improved through mentorship, professional development, and participation in regional and international REDCap Consortium activities.

Textbox 2. Summary of support strategies for electronic data capture (EDC) adoption at the University of the Witwatersrand Faculty of Health Sciences.

Support strategy and observations and effects

- Top-down organizational support
  - Official endorsement from management ensures that an EDC system receives adequate resources for infrastructure and staff and signals institutional commitment to potential end users.
- Secure and reliable application, hosting infrastructure, and systems administration
  - Prioritizing reliability and security of the EDC system builds trust among users. In resource-constrained settings, power, network, and information infrastructure is often unreliable, and users fear losing their data.
- An enabling and accessible REDCap (Research Electronic Data Capture) support team
  - The availability of a person or team acting as technology bridgers [31] reduces anxiety and resistance associated with technology adoption.
- Regular hands-on training workshops
  - Structured practical training opportunities capacitate new users with an EDC system and serve as a mechanism to disseminate knowledge and best practices.
- Annual conferences or symposia for end users
  - Regular academic events that promote the correct use of features and demonstrate the benefits of an EDC system attract new users while keeping existing users informed and engaged.
- Participation in international REDCap Consortium activities
  - REDCap administrators' knowledge and abilities are developed through interaction with peers from other international institutions.
- Mentorship- and capacity-development relationships with established organizations
  - Institutions that do not have established clinical research informatics departments or capacity benefit greatly from mentorship and collaboration with experienced partner institutions.

Gap Analysis

Although adoption has been a success, there were gaps in our processes, which are important to recognize, especially for those starting the process. Although top-down organizational support has been strong, funding has always been difficult in a resource-constrained environment. A user-pays model can be attractive, but implementation is difficult and may encourage users to stick to paper and spreadsheets. Funding from the center may be politically fraught, especially in the early phases before the system has proved itself. We have gained stability over the last 10 years but are still working on improving the financial model. A related issue is the size of the support team—dedicated staff are required, and in resource-constrained environments this may be hard to find or pay for. In our case, we were fortunate to have dedicated staff members from the start, but they were always working under pressure, which limited the extent of training and ability to support strategic projects. This latter issue was mitigated to some extent by the strong mentoring role played by the VUMC. Investment in the support team through attendance at international events has been very important, but resource constraints have limited how many individuals can attend and how frequently.

Limitations of This Study

The principal investigator on this study (IM) is also the lead REDCap administrator for the Wits FHS. This makes them intimately familiar with the support processes and end-user interactions at the institution but can lead to a lack of objectivity. For this reason, coauthors experienced with REDCap and clinical research informatics from outside of our institution were included to provide a more balanced and fair report.

Conclusions

The implementation of technology requires strategies to support, and manage resistance from, end users. One of the main reasons for this resistance is inertia: end users naturally resist change to familiar and deeply ingrained processes [1,5,7,15-17]. In our experience, and supported by findings from others [1,4,7,16,17], to overcome inertia, one needs to demonstrate the reliability of the system and benefits of adoption to prospective users, while at the same time easing the transition process by providing adequate end-user support. The capacity to implement and support REDCap at the Wits FHS was initiated through organizational and financial backing of the FHS management. This capacity was subsequently developed further through participation in REDCap Consortium activities such as REDCapCon and the REDCap Community forum and through a strong bidirectional relationship with the VUMC, the institution that created REDCap.

Acknowledgments

This research was conducted as part of a postgraduate qualification (Master of Science in Medicine: Biomedical Informatics and Translational Medicine) by the first author, IAM. Funding support for open access publishing is provided by the Department of Surgery at the University of the Witwatersrand Faculty of Health Sciences. The authors wish to acknowledge Michelle Jones and Ndzalama Shivambu, who contributed to the review of publishing and bibliometric data reported in this paper.

Authors' Contributions

The project was conceptualized by MK, IAM, PAH, and RZ. Most of the writing was done by IAM and MK, with sections contributed by SH, BK, RZ, and PAH. IAM and MDN collected, analyzed, and prepared the data about REDCap system use, training and symposia attendance, and bibliometric data. BK, RZ, MDN, SH, and PAH contributed to editing and review of the manuscript.

Conflicts of Interest

None declared.

References

1. Anderson NR, Lee ES, Brockenbrough JS, Minie ME, Fuller S, Brinkley J, et al. Issues in biomedical research data management and analysis: needs and barriers. J Am Med Inform Assoc 2007;14(4):478-488 [FREE Full text] [doi: 10.1197/jamia.M2114] [Medline: 17460139]
2. Welker JA. Implementation of electronic data capture systems: barriers and solutions. Contemp Clin Trials 2007 May;28(3):329-336. [doi: 10.1016/j.cct.2007.01.001] [Medline: 17287151]
3. Castle C. Converting from Paper-Based to Electronic Data Capture and Record Keeping in Clinical Trial Management: Benefits, Challenges and Practical Considerations. University of North Texas. 2015 Nov. URL: https://unthsc-ir.tdl.org/bitstream/handle/20.500.12503/28998/2015_12_gsbs_Castle_Colton_practicum.pdf?sequence=1&isAllowed=y [accessed 2016-07-31]
4. Landis-Lewis Z, Manjomo R, Gadabu OJ, Kam M, Simwaka BN, Zickmund SL, et al. Barriers to using eHealth data for clinical performance feedback in Malawi: a case study. Int J Med Inform 2015 Oct;84(10):868-875 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.07.003] [Medline: 26238704]
5. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
6. Furusa SS, Coleman A. Factors influencing e-health implementation by medical doctors in public hospitals in Zimbabwe. S Afr J Inf Manag 2018 Jun 14;20(1):a928. [doi: 10.4102/sajim.v20i1.928]
7. Schopf TR, Nedrebø B, Hufthammer KO, Daphu IK, Lærum H. How well is the electronic health record supporting the clinical tasks of hospital physicians? A survey of physicians at three Norwegian hospitals. BMC Health Serv Res 2019 Dec 04;19(1):934 [FREE Full text] [doi: 10.1186/s12913-019-4763-0] [Medline: 31801518]
8. Klipin M, Mare I, Hazelhurst S, Kramer B. The process of installing REDCap, a Web based database supporting biomedical research: the first year. Appl Clin Inform 2014 Nov 19;5(4):916-929 [FREE Full text] [doi: 10.4338/ACI-2014-06-CR-0054] [Medline: 25589907]
9. Ash JS, Anderson NR, Tarczy-Hornoch P. People and organizational issues in research systems implementation. J Am Med Inform Assoc 2008;15(3):283-289 [FREE Full text] [doi: 10.1197/jamia.M2582] [Medline: 18308986]
10. Murphy SN, Dubey A, Embi PJ, Harris PA, Richter BG, Turisco F, et al. Current state of information technologies for the clinical research enterprise across academic medical centers. Clin Transl Sci 2012 Jun;5(3):281-284 [FREE Full text] [doi: 10.1111/j.1752-8062.2011.00387.x] [Medline: 22686207]
11. Bukachi F, Pakenham-Walsh N. Information technology for health in developing countries. Chest 2007 Nov;132(5):1624-1630. [doi: 10.1378/chest.07-1760] [Medline: 17998362]
12. Yogeswaran P, Wright G. EHR implementation in South Africa: how do we get it right? Stud Health Technol Inform 2010;160(Pt 1):396-400. [Medline: 20841716]
13. McLean E, Dube A, Saul J, Branson K, Luhanga M, Mwiba O, et al. Implementing electronic data capture at a well-established health and demographic surveillance site in rural northern Malawi. Glob Health Action 2017;10(1):1367162 [FREE Full text] [doi: 10.1080/16549716.2017.1367162] [Medline: 28922071]
14. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
15. Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. J Investig Med 2005 May;53(4):192-200. [doi: 10.2310/6650.2005.00402] [Medline: 15974245]

16. Grossman M, Bates S. Knowledge capture within the biopharmaceutical clinical trials environment. VINE 2008;38(1):118-132. [doi: 10.1108/03055720810870932]
17. Franklin JD, Guidry A, Brinkley JF. A partnership approach for Electronic Data Capture in small-scale clinical trials. J Biomed Inform 2011 Dec;44 Suppl 1:S103-S108 [FREE Full text] [doi: 10.1016/j.jbi.2011.05.008] [Medline: 21651992]
18. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc 2009;16(3):316-327 [FREE Full text] [doi: 10.1197/jamia.M3005] [Medline: 19261934]
19. South Africa | Data. The World Bank. 2021. URL: https://data.worldbank.org/country/south-africa [accessed 2021-05-20]
20. Coovadia H, Jewkes R, Barron P, Sanders D, McIntyre D. The health and health system of South Africa: historical roots of current public health challenges. Lancet 2009 Sep 05;374(9692):817-834. [doi: 10.1016/S0140-6736(09)60951-X] [Medline: 19709728]
21. Bradshaw D, Nannan NN, Pillay-van Wyk V, Laubscher R, Groenewald P, Dorrington RE. Burden of disease in South Africa: protracted transitions driven by social pathologies. S Afr Med J 2019 Dec 05;109(11b):69-76. [doi: 10.7196/SAMJ.2019.v109i11b.14273] [Medline: 32252872]
22. Facts and figures - Wits University. University of the Witwatersrand, Johannesburg. 2020. URL: https://www.wits.ac.za/about-wits/facts-and-figures/ [accessed 2021-05-24]
23. About Us - Wits University. University of the Witwatersrand, Johannesburg. 2020. URL: http://www.wits.ac.za/health/about-us/ [accessed 2021-06-19]
24. Kramer B, Libhaber E. Closing the barrier between disease and health outcomes in Africa through research and capacity development. Glob Health Action 2018;11(1):1425597 [FREE Full text] [doi: 10.1080/16549716.2018.1425597] [Medline: 29370732]
25. 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]
26. Kramer B, Zent R. Diaspora linkages benefit both sides: a single partnership experience. Glob Health Action 2019;12(1):1645558 [FREE Full text] [doi: 10.1080/16549716.2019.1645558] [Medline: 31362603]
27. About Nagios. What is Nagios? Nagios Enterprises. URL: https://www.nagios.org/about/ [accessed 2022-02-15]
28. Teraco Data Environments. Teraco. URL: https://www.teraco.co.za/ [accessed 2021-09-11]
29. Wits Health Consortium - Home. Wits Health Consortium. 2021. URL: https://www.witshealth.co.za/ [accessed 2021-09-11]
30. Arcserve | Data Protection and Business Continuity Solutions. Arcserve. URL: https://www.arcserve.com/node/1 [accessed 2022-02-15]
31. Ash JS, Stavri PZ, Dykstra R, Fournier L. Implementing computerized physician order entry: the importance of special people. Int J Med Inform 2003 Mar;69(2-3):235-250. [doi: 10.1016/s1386-5056(02)00107-7] [Medline: 12810127]
32. Edmondson AC. The Fearless Organization: Creating Psychological Safety in the Workplace for Learning, Innovation, and Growth. Hoboken, NJ, USA: John Wiley & Sons; 2018.
33. Nadan CH. Software licensing in the 21st century: are software licenses really sales, and how will the software industry respond. AIPLA Q J 2004;32(4):555.
34. Partners. REDCap. URL: https://projectredcap.org/partners/ [accessed 2017-08-15]
35. REDCap Community Site. REDCap. 2021. URL: https://projectredcap.org/resources/community/ [accessed 2021-05-16]
36. Hofman K, Kramer B. Human resources for research: building bridges through the Diaspora. Glob Health Action 2015 Nov 6;8:29559 [FREE Full text] [doi: 10.3402/gha.v8.29559] [Medline: 26548635]
37. Taylor R. MySQL Simple Admin | Vanderbilt University Medical Center. GitHub. 2021. URL: https://github.com/vanderbilt-redcap/mysql-simple-admin#mysql-simple-admin [accessed 2021-01-08]
38. Accessible Law – Protection of Personal Information Act (POPI Act). Protection of Personal Information Act. 2020. URL: https://popia.co.za/ [accessed 2021-12-24]
39. Pang X, Kozlowski N, Wu S, Jiang M, Huang Y, Mao P, et al. Construction and management of ARDS/sepsis registry with REDCap. J Thorac Dis 2014 Sep;6(9):1293-1299 [FREE Full text] [doi: 10.3978/j.issn.2072-1439.2014.09.07] [Medline: 25276372]
40. Tuti T, Bitok M, Paton C, Makone B, Malla L, Muinga N, et al. Innovating to enhance clinical data management using non-commercial and open source solutions across a multi-center network supporting inpatient pediatric care and research in Kenya. J Am Med Inform Assoc 2016 Jan;23(1):184-192 [FREE Full text] [doi: 10.1093/jamia/ocv028] [Medline: 26063746]
41. Campion TR, Pompea ST, Turner SP, Sholle ET, Cole CL, Kaushal R. A method for integrating healthcare provider organization and research sponsor systems and workflows to support large-scale studies. AMIA Jt Summits Transl Sci Proc 2019 May 6;2019:648-655 [FREE Full text] [Medline: 31259020]
42. Gesell SB, Halladay JR, Mettam LH, Sissine ME, Staplefoote-Boynton BL, Duncan PW. Using REDCap to track stakeholder engagement: a time-saving tool for PCORI-funded studies. J Clin Transl Sci 2020 Apr;4(2):108-114 [FREE Full text] [doi: 10.1017/cts.2019.444] [Medline: 32313700]

Abbreviations

EDC: electronic data capture
FTE: full-time equivalent
POPIA: Protection of Personal Information Act
REDCap: Research Electronic Data Capture
REDCapCon: Research Electronic Data Capture conference
VUMC: Vanderbilt University Medical Center
Wits FHS: University of the Witwatersrand Faculty of Health Sciences

Edited by C Lovis; submitted 14.10.21; peer-reviewed by C Gaudet-Blavignac, Y Chu; comments to author 02.01.22; revised version received 01.03.22; accepted 31.05.22; published 30.08.22.

Please cite as:
Maré IA, Kramer B, Hazelhurst S, Nhlapho MD, Zent R, Harris PA, Klipin M
Electronic Data Capture System (REDCap) for Health Care Research and Training in a Resource-Constrained Environment: Technology Adoption Case Study
JMIR Med Inform 2022;10(8):e33402
URL: https://medinform.jmir.org/2022/8/e33402
doi:10.2196/33402
PMID:36040763

Review

Uncertainty Estimation in Medical Image Classification: Systematic Review

Alexander Kurz[1], MSc; Katja Hauser[1], MSc; Hendrik Alexander Mehrtens[1], MSc; Eva Krieghoff-Henning[1], PhD; Achim Hekler[1], MSc; Jakob Nikolas Kather[2], Prof Dr med; Stefan Fröhling[3], Prof Dr med; Christof von Kalle[4], Prof Dr med; Titus Josef Brinker[1], Dr med

[1]Digital Biomarkers for Oncology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany

[3]Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany

[4]Department of Clinical-Translational Sciences, Berlin Institute of Health (BIH), Berlin, Germany

Corresponding Author:
Titus Josef Brinker, Dr med
Digital Biomarkers for Oncology Group
German Cancer Research Center (DKFZ)
Im Neuenheimer Feld 280
Heidelberg, 69120
Germany
Phone: 49 62213219304
Email: titus.brinker@nct-heidelberg.de

Abstract

Background: Deep neural networks are showing impressive results in different medical image classification tasks. However, for real-world applications, there is a need to estimate the network's uncertainty together with its prediction.

Objective: In this review, we investigate in what form uncertainty estimation has been applied to the task of medical image classification. We also investigate which metrics are used to describe the effectiveness of the applied uncertainty estimation

Methods: Google Scholar, PubMed, IEEE Xplore, and ScienceDirect were screened for peer-reviewed studies, published between 2016 and 2021, that deal with uncertainty estimation in medical image classification. The search terms "uncertainty," "uncertainty estimation," "network calibration," and "out-of-distribution detection" were used in combination with the terms "medical images," "medical image analysis," and "medical image classification."

Results: A total of 22 papers were chosen for detailed analysis through the systematic review process. This paper provides a table for a systematic comparison of the included works with respect to the applied method for estimating the uncertainty.

Conclusions: The applied methods for estimating uncertainties are diverse, but the sampling-based methods Monte-Carlo Dropout and Deep Ensembles are used most frequently. We concluded that future works can investigate the benefits of uncertainty estimation in collaborative settings of artificial intelligence systems and human experts.

International Registered Report Identifier (IRRID): RR2-10.2196/11936

(JMIR Med Inform 2022;10(8):e36427) doi:10.2196/36427

KEYWORDS

uncertainty estimation; network calibration; out-of-distribution detection; medical image classification; deep learning; medical imaging

Introduction

Overview

Digital image analysis is a helpful tool to support physicians in their clinical decision-making. Originally, digital image analysis was performed by extracting handcrafted features from an input image. These features can be tuned to the underlying data, which means that for a specific disease, only specific features can be looked for in the observed image. With the advent of deep learning, however, a "black box" has been established that can, in the setting of supervised learning, intrinsically learn such

features from labeled data. In recent years, deep learning–based methods have vastly outperformed traditional methods that rely on handcrafted features. With the learning-based methods, the focus has shifted from manually defining image features to providing clean and correctly annotated data to the learning system. With the data-centric approach, however, new challenges arise.

In a clinical setting, when such algorithms are meant to be used as diagnostic assistance tools, the user has to be able to understand how the artificial intelligence (AI) system came up with the diagnosis. One key component in this regard is a measure of confidence of the AI system in its prediction. Such a measure is important to increase trust in the AI system, and it may improve clinical decision-making [1]. We will use the term "uncertainty estimation" for measures to evaluate model confidence. When the AI system provides a measure for its uncertainty, predictions with high uncertainties can be treated with extra care by medical experts. On the other hand, the human expert can better trust the prediction of an AI system where it reports low uncertainty. In this study, we review recent publications that have applied uncertainty estimation methods to medical image classification tasks. The area of uncertainty estimation in deep neural networks is an active research field, and the currently most popular methods have been proposed from 2016 onward. In the next section, we provide an overview of the most prominent methods for uncertainty estimation.

In the results section, we categorize the reviewed works by the uncertainty estimation method they apply. We provide a table that serves as an overview of all the included studies. In the last section, we discuss the most frequently used metrics for evaluating the benefit of uncertainty estimation and give an outlook of possible future research directions with a focus on human-machine collaboration.

Technical Background

In a classification task, the neural network is supposed to predict how likely it is for a given input x to belong to class y out of a fixed number of possible classes. The output of the neural network can be interpreted as a probability distribution over all classes, with each individual value indicating how likely it is for the input to belong to the respective class.

In formula, the predictive distribution can be written as follows:

The predictive distribution given input x and training data D is described as the integral over the likelihood $p(y|x,\theta)$ with prior $p(\theta|D)$ computed over the model's parameters θ. In deep neural networks, this integral cannot be computed analytically. Therefore, methods that try to capture uncertainty in neural networks try to approximate the predictive distribution.

Depending on the modeled uncertainty, the predictive uncertainty can be divided into aleatoric uncertainty and epistemic uncertainty. The aleatoric uncertainty describes the uncertainty inherent in the data, whereas the epistemic uncertainty captures the uncertainty of the model. The softmax output of a typical classification network is only able to capture aleatoric uncertainty [2].

Methods for Uncertainty Estimation

Ovadia et al [3] compared several popular methods for uncertainty estimation. In this work, we name the methods that we discovered to be most popular and refer the reader to the respective works for a detailed description of the proposed methods. We categorize the methods into (1) model sampling, (2) single network methods, and (c) data augmentation.

Model Sampling

Sampling-based methods are easy to implement as they make use of existing network architectures. The 2 most popular methods are Monte Carlo dropout (MCDO) [4] and Deep Ensembles [5]. Both methods rely on several prediction runs of either an ensemble of multiple neural networks or a neural network with dropout layers to compute a predictive uncertainty.

Single Network Methods

The field of directly modifying the network architecture for improved uncertainty estimation is quite diverse. In the derivation of MCDO, the authors compare their approach to Gaussian processes (GPs). A GP is a method to estimate a distribution over functions [6] and can be applied to estimate uncertainties in neural networks.

Approaches that have been included in the comparison by Ovadia et al [3] include stochastic variational inference (SVI) [7] and temperature scaling (TS) [8]. SVI applies the concept of variational inference to deep neural networks, whereas TS works as a post hoc method. By applying a scaling factor to the network output, TS can improve network calibration. Another method worth mentioning is evidential deep learning (EDL) [9]. EDL fits a Dirichlet distribution to the network output to estimate the network's uncertainty.

Data Augmentation

Comparable to sampling multiple models, one can also compute a distribution of predictions by running the network on different augmentations of the input data. Ayhan and Berens [10] propose such a method for improved aleatoric uncertainty estimation called test-time data augmentation (TTA).

Methods

Data Extraction

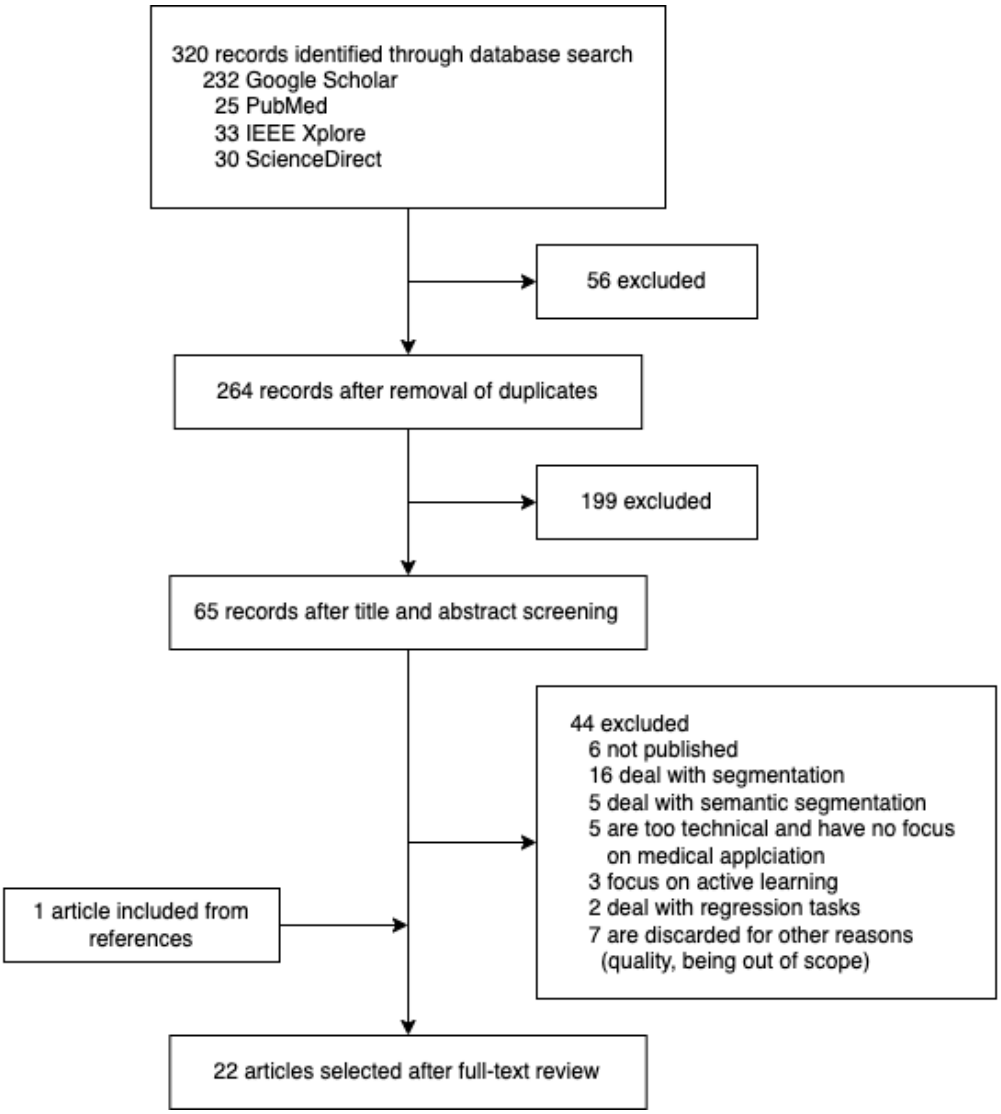
For the systematic review, we searched through Google Scholar, PubMed, IEEE Xplore, and ScienceDirect to identify relevant works that apply uncertainty estimation methods to medical image classification. We limited our search to works that have appeared between January 2016 and October 2021. As search terms, we used "uncertainty," "uncertainty estimation," "network calibration," and "out-of-distribution detection," and we combined them with the terms "medical images," "medical image analysis," and "medical image classification."

Selection Process

The selection process was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [11]. We found 320 potentially relevant publications from the database search. During title and abstract screening, we discarded the majority of the works, as

they either did not estimate uncertainties at all or dealt with other image analysis problems such as image segmentation. From the first screening round, 65 papers were selected for full-text analysis. During the full-text analysis, we discarded several other works, as they turned out to deal with other problems including semantic segmentation. Eventually, 22 papers were included in the review. Figure 1 visualizes the selection process.

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

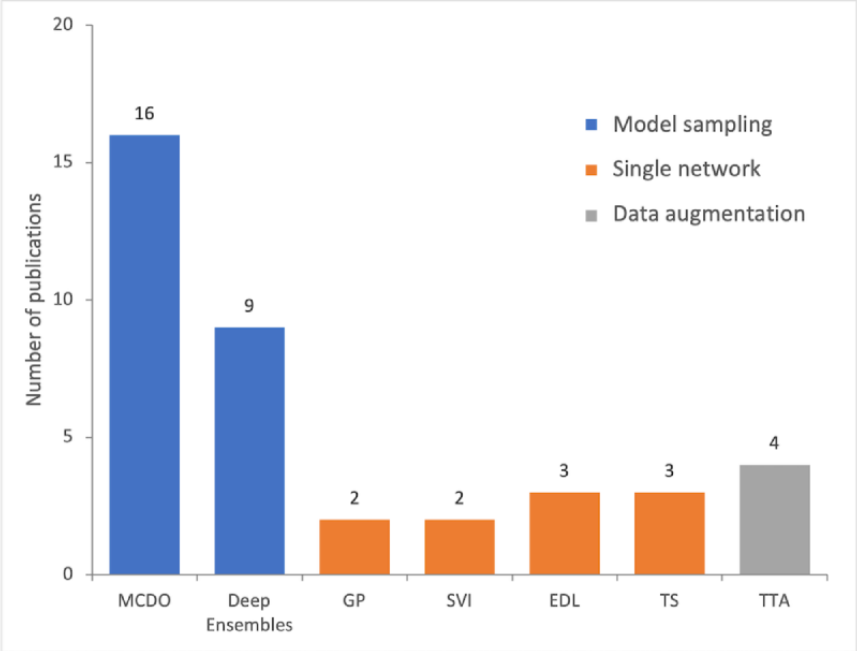


Results

Paper Categorization

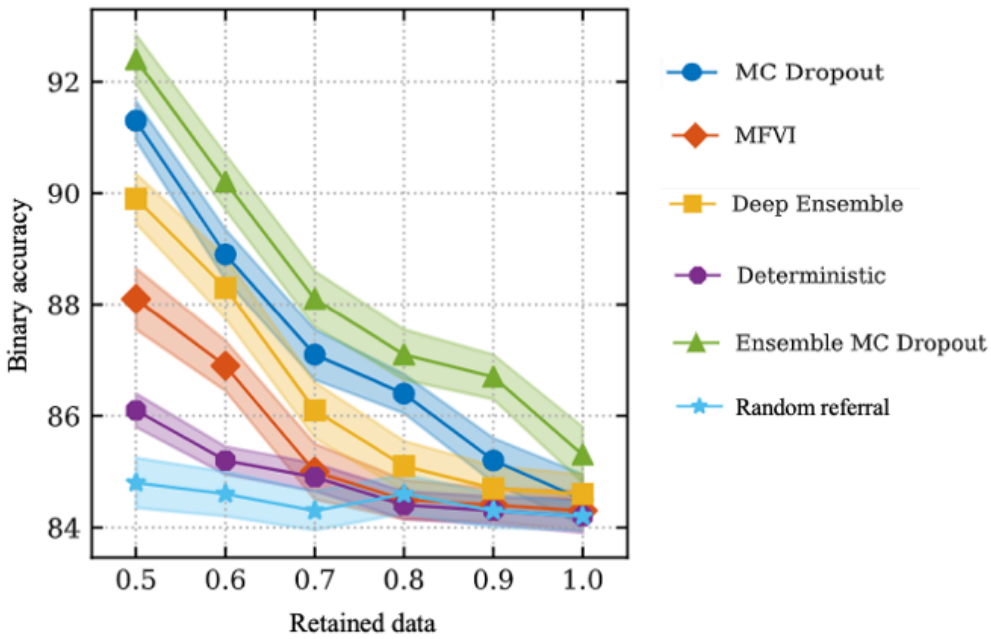
Figure 2 provides an overview of the applied methods in all of the reviewed works. Note that most included works apply more than 1 method for uncertainty estimation. We observed that the majority of works apply sampling-based methods (ie, MCDO and Deep Ensembles). In the category that we denoted as single network methods, all corresponding methods are almost equally represented. Lastly, 4 works that we included apply TTA to compute an uncertainty estimate.

Figure 2. Number of publications that apply the respective uncertainty estimation method. EDL: evidential deep learning; GP: Gaussian process; MCDO: Monte Carlo dropout; SVI: stochastic variational inference; TS: temperature scaling; TTA: test-time data augmentation.



Most of the included works evaluate the applied methods by computing an uncertainty measure (mostly predictive variance or predictive entropy). This uncertainty measure is often used to generate retained data versus accuracy evaluations. Figure 3 shows an example of retained data versus accuracy plot from the study by Filos et al [2]. From the plot, it can be observed that when only the more certain samples are retained, accuracy on the retained data increases. The methods for uncertainty estimation are then ranked by how far they increase the accuracy on the retained data.

Figure 3. Retained data versus accuracy plot from Filos et al [2]. MFVI: mean field variational inference.



Some included works focus on network calibration and try to decrease the expected calibration error (ECE) within their experiments. Some other works use the computed uncertainty measure to detect out-of-distribution (OOD) samples. Table 1 provides an overview of all included works. In the following sections, we will briefly cover the content of each included study.

Table 1. Overview of the selected studies.

Methods	Organs or sickness	Sensor	Network architecture	Reported metrics	Data access	Code available	Reference
MCDO[a], GP[b]	Diabetic retinopathy from fundus images	Camera	Custom CNNs[c]	Retained data or accuracy, uncertainty or density	Public (Kaggle competition)	Yes	Leibig et al [12]
MCDO, SVI[d]	Retina	Optical coherence tomography	ResNet-18	Predictive variance	Public	Yes	Laves et al [13]
MCDO	Skin cancer	Camera	VGG-16, ResNet-50, DenseNet-169	Uncertainty or density, retained data or accuracy, uncertainty, confusion matrix	Public	Yes	Mobiny et al [14]
MCDO	Brain	MRI[e]	Modified VGGNet	Reliability diagrams, AUROC[f]	Private	Yes	Herzog et al [15]
MCDO	Breast cancer	Mammography	VGG-19	Uncertainty, confusion matrix	Public	No	Caldéron-Ramírez et al [16]
MCDO, DUQ[g]	COVID-19	X-ray	WideResNet	Jensen-Shannon divergence	Public	No	Caldéron-Ramírez et al [17]
MCDO, Ensembles, MFVI[h]	Diabetic retinopathy from fundus images	Camera	VGG Variants	Retained data or accuracy, retained data or AUROC, ROC[i]	Public (Kaggle competition)	Yes	Filos et al [2]
MCDO, Ensembles, M-heads	Histopathological slides	Microscope	DenseNet	Retained data or AUROC	Public	No	Linmans et al [18]
MCDO, Ensembles, Mix-up	Histopathological slides	Microscope	ResNet-50	ECE[j], AUROC, AUPRC[k]	Private	No	Thagaard et al [19]
MCDO, Ensembles	COVID-19, Histopathological slides (breast cancer)	CT[l], microscope	ResNet-152-V2, Inception-V3, Inception-ResNet-V2	Predictive entropy, retained data or accuracy	Public	No	Yang and Fevens [20]
MCDO, Ensembles, TWD[m]	Skin cancer	Camera	ResNet-152, Inception- ResNet-V2, DenseNet-201, MobileNet-V2	Entropy, AUROC	Public (Kaggle competition, ISIC data set)	No	Abdar et al [21]
MCDO, Ensembles, others	Lung	X-ray	WideResNet	AUROC, AUPRC	Public	No	Berger et al [22]
GP	Diabetic retinopathy from fundus images	Camera	Inception-V3	AUROC	Public (Kaggle competition)	Yes	Toledo-Cortés et al [23]
EDL[n] + Ensembles	Chest	X-ray	DenseNet-121	AUROC	Public	No	Ghesu et al [24]
EDL + MCDO	Breast cancer	Mammography	VGGNet	AUROC	Public + private	No	Tardy et al [25]
EDL	Chest, abdomen, and brain	X-ray, ultrasound, MRI	DenseNet-121	AUROC, coverage or F1 score, coverage or AUROC	Public	No	Ghesu et al [26]
TS[o], MCDO	Polyp	Colonoscopy (camera)	ResNet-101, DenseNet-121	ECE, predictive entropy, predictive variance	Public + private	No	Carneiro et al [27]
TS, DCA[p]	Head CT, mammography, chest x-ray, histology	Multimodal	AlexNet, ResNet-50, DenseNet-121, SqueezeNet	ECE	Public	No	Liang et al [28]

Methods	Organs or sickness	Sensor	Network architecture	Reported metrics	Data access	Code available	Reference
TTA[q]	Diabetic retinopathy from fundus images	Camera	ResNet-50	Uncertainty or density, retained data or AUROC	Public (Kaggle competition)	Yes	Ayhan and Berens [10]
TTA, MCDO, MCBN[r], Ensembles	Skin cancer	Camera	ResNet-50	ECE	Private (31,000 annotated images)	No	Jensen et al [29]
TTA + MCDO	Skin cancer	Camera	Efficient-Net-B0	Predictive entropy, predictive variance, Bhattacharya coefficient, retained data or accuracy	Public (ISIC data set)	No	Combalia et al [30]
TTA, TS, Ensembles	Diabetic retinopathy from fundus images	Camera	Modified ResNet	Reliability diagrams, AECE[s], retained data or AUROC	Public (Kaggle competition)	Yes	Ayhan et al [31]

[a]MCDO: Monte Carlo dropout.
[b]GP: Gaussian process.
[c]CNN: convolutional neural network.
[d]SVI: stochastic variational inference.
[e]MRI: magnetic resonance imaging.
[f]AUROC: area under the receiver operating curve.
[g]DUQ: deterministic uncertainty quantification.
[h]MFVI: mean field variational inference.
[i]ROC: receiver operating curve.
[j]ECE: expected calibration error.
[k]AUPRC: area under the precision recall curve.
[l]CT: computed tomography.
[m]TWD: three-way decision theory.
[n]EDL: evidential deep learning.
[o]TS: temperature scaling.
[p]DCA: difference between confidence and accuracy.
[q]TTA: test-time data augmentation.
[r]MCBN: Monte-Carlo batch norm.
[s]AECE: adaptive expected calibration error.

Sampling-Based Methods

The first work that we have included is the study by Leibig et al [12], which applies MCDO to the task of diabetic retinopathy classification. To evaluate the impact of the applied uncertainty estimation method, the authors report retained data versus accuracy curves. This means that a fraction of uncertain predictions is discarded, and it is evaluated how discarding uncertain samples can improve the accuracy on the test data set. The results show that discarding 20% or more of the most uncertain samples can notably improve the accuracy of the neural network. In their work, the authors compare the performance of MCDO to an alternatively implemented GP and find that MCDO leads to better accuracies on the retained data versus accuracy evaluations.

Laves et al [13] apply MCDO and SVI to retina scans observed through optical coherence tomography. The authors show that both methods lead to higher standard deviations on false-positive predictions compared to true positive predictions. This indicates that the standard deviations can be used to refer predictions with high uncertainty to human experts to improve the classification accuracy.

Mobiny et al [14] estimate uncertainties using MCDO with different types of networks including VGGNet [32], ResNet [33], and DenseNet [34] on dermoscopic images of 8 different skin lesion types. Similar to Leibig et al [12], the authors report retained data versus accuracy curves and show that the accuracy can be increased when referring a fraction of uncertain samples to a medical expert. As a measure for uncertainty, the normalized predictive entropy is computed. As an additional metric, the authors also compute an uncertainty-related confusion matrix that includes the numbers of correct-certain, correct-uncertain, incorrect-certain, and incorrect-uncertain predictions. The respective numbers vary when the uncertainty threshold is changed. One possible goal with this evaluation is to decrease the number of incorrect-certain predictions as much as possible.

Another work by Herzog et al [15] applies MCDO to the classification of brain magnetic resonance imaging (MRI) images. The goal of their work is to infer patient-level diagnostics from the predictions from multiple images. Therefore, the authors compute a variety of 5 uncertainty measures per image. To draw conclusions on a patient level, the authors run another neural network that processes the uncertainties of all images belonging to one patient.

In two other published works, Caldéron-Ramírez et al [16,17] apply MCDO to the tasks of breast cancer classification from mammography images and to COVID-19 classification from chest x-ray scans. Unfortunately, even among the two works, the authors report different metrics, which prevents comparing the results. In the breast cancer classification task, the authors use a metric called uncertainty balanced accuracy, which builds up on the uncertainty-related confusion matrix also used by Mobiny et al [14]. In the work related to COVID-19 detection, the authors report the Jensen-Shannon divergence as an uncertainty measure, which we did not encounter in any of the other reviewed works.

Another set of studies compared MCDO to Deep Ensembles (further simply denoted as Ensembles) and partly to other methods. Filos et al [2] compare MCDO to Ensembles and mean field variational inference (MFVI), which is a variation of SVI, and apply it to the task of diabetic retinopathy classification. In addition to comparing MCDO and Ensembles individually, they also combine both approaches and include the combination in the evaluation, denoted as "Ensemble MCDO." As neural network architecture, the authors use variants of VGGNet [32]. The retained data versus accuracy plots show that "Ensemble MCDO" leads to the best performance, followed by MCDO and Ensembles applied individually. MFVI did not achieve the same performance as the sampling-based methods.

Linmans et al [18] perform uncertainty estimation on the publicly available Camelyon data sets for breast cancer detection on histopathological slides. The authors propose a new method for uncertainty estimation called "M-heads," which adds multiple output heads to the convolutional neural network (CNN). They compare their method to the MCDO and Ensembles of 5 and 10 networks, respectively. From the different evaluations, the confidence versus accuracy plot shows that accuracy increases when only keeping predictions with high confidence. The methods rank from M-heads performing best, followed by the Ensembles of 5 and 10 networks. In the reported results, MCDO does not perform better than the vanilla softmax output.

Thagaard et al [19] apply Ensembles and MCDO to private data sets of histopathological slides for breast cancer detection. In their work, the authors focus on OOD detection while analyzing combinations of different internal data sets. Concerning the comparison of the uncertainty estimation methods, the ECE is calculated on 3 different data sets. For all 3 data sets, the Ensemble of 5 ResNet-50 networks reaches the best ECE scores.

In another work, Yang and Fevens [20] apply MCDO, Ensembles, and a combination of both to several publicly available data sets. The modalities include COVID-19 classification from x-ray scans, brain tumor classification from MRI images, and breast cancer detection from histopathological slides. On the histopathological images, the authors present retained data versus accuracy plots. For the reported accuracies, the Ensemble MCDO approach with 5 Inception-ResNet networks leads to the best results.

Abdar et al [21] apply MCDO, Ensembles, and Ensemble MCDO to skin cancer classification from dermoscopic images. The authors report entropies and standard deviations of the applied methods for 4 different network architectures on 2 different publicly available data sets. From the reported values, the authors conclude that the Ensembles overall perform best. In an additional setup, the authors combine 2 uncertainty estimation methods (Ensembles and Ensembles MCDO) in a decision tree that they refer to as 3-way decision theory.

In another work, Berger et al [22] evaluate confidence-based OOD detection on x-ray scans of lung diseases. The authors compare MCDO, Ensembles, and specific methods for OOD detection, including a method based on Mahalanobis distance and the "out-of-distribution detector for neural networks" [35]. In their experiments, the authors find that the OOD detector for neural networks leads to the best results for OOD detection with respect to the area under the receiver operating curve (AUROC) and area under the precision recall curve (AUPRC) values.

Single Network Methods

After having covered several works that focus on sampling-based uncertainty estimation methods, we will now look into works that directly apply to the network's classification output to estimate uncertainties. One example is the work by Toledo-Cortés et al [23] that applies a GP to the output of their implemented Inception-V3 network [36]. Similar to Laves et al [13], the authors report standard deviations on true positive and false positive predictions. Since the standard deviations for both cases are quite similar, it must be concluded that the applied GP is not well suited for a useful uncertainty estimation.

A set of other works applies EDL to estimate uncertainties. In their first work, Ghesu et al [24] work with x-ray scans of the chest and later extend their approach to ultrasound images of the abdomen and MRI images of the brain [26]. The results show that discarding a fraction of the most uncertain predictions can notably improve the AUROC score averaged over different x-ray classification tasks.

Comparably, Tardy et al [25] apply EDL to the task of breast cancer classification from mammography images. The authors also report improved AUROC and AUPRC values when discarding a fraction of uncertain samples.

Two works that we have included apply TS to medical image classification tasks. Carneiro et al [27] combine TS and MCDO to compute a calibrated confidence measure as well as an uncertainty measure in the form of predictive entropy and predictive variance. The authors evaluate the methods on 2 different cohorts of colonoscopy images with respect to a 5-class polyp classification task. The reported ECE and accuracy values show that the DenseNet-121 architecture with both MCDO and TS leads to the best results.

Liang et al [28] present a new approach for network calibration in the form of an auxiliary loss term called "difference between confidence and accuracy" (DCA) that can be integrated into an existing CNN training procedure. The authors compare their approach to TS and uncalibrated networks on different medical data sets with several different network architectures. The results show that in most cases, DCA produces the best ECE values. It is also shown that depending on the data set and model architecture, TS does not always improve the expected calibration error.

Test-Time Data Augmentation (TTA)

The concept of TTA is introduced by Ayhan and Behrens [10], where it is applied to the task of diabetic retinopathy from fundus images. The authors apply 128 different augmentations, ranging from cropping and resizing to different color augmentations. As measure for uncertainty, the interquartile range of the predictions is computed. Similar to Leibig et al [12], the authors report retained data versus AUROC curves and show that the AUROC values improve when referring uncertain samples to a medical expert.

Another work by Jensen et al [29] focuses on evaluating interrater agreement on dermoscopic images of different skin lesions. In the experiment, multiple experts have provided labels for the respective images, and the labels for each sample can vary across experts. Therefore, the approaches of label fusion and label sampling are compared for training the neural network. These approaches are combined with methods that estimate uncertainties to evaluate the influence on the network's calibration of the combined methods. It is shown that in the specific experimental setting, the combination of label sampling and TTA leads to the highest classification accuracies among all data splits.

Combalia et al [30], also working with dermoscopic images, combine TTA and MCDO to evaluate aleatoric as well as epistemic uncertainties. In their experiments, the authors show that the combination of both methods leads to the best results for OOD detection as well as on the retained data versus accuracy evaluation. For the evaluations, 100 forward passes through the network are performed with either TTA or MCDO or both methods combined. The uncertainties are quantified by computing the predictive entropy, the predictive variance, and additionally, the Bhattacharyya coefficient [30].

In a follow-up of their original work, Ayhan et al [31] extend their experiments on diabetic retinopathy classification by other uncertainty estimation methods. Besides TTA, the authors also include TS and an ensemble of 3 modified ResNet networks. To compare the results, the authors compute The Adaptive Expected Calibration Error [37]. In terms of Adaptive Expected Calibration Error, the median probability of 128 forward passes with different data augmentations leads to the best calibrated results. On the retained data versus AUROC curves, TTA and Deep Ensembles perform equally well. The experiments on a different cohort of fundus images show that TS generalizes worse to new data compared to TTA and Deep Ensembles.

Discussion

Through the reviewed publications, we gained an overview of which methods for uncertainty estimation are most frequently used in the field of medical image classification. We found that the sampling-based methods MCDO and Deep Ensembles are the most frequently applied methods. With the sampling-based approaches, it is possible to compute a distribution of predictions and from there determine an uncertainty measure, usually either in the form of predictive entropy or predictive variance. These measures help to identify samples where the neural network is uncertain about its predictions.

In addition to the sampling-based uncertainty evaluations, we also observed evaluations that analyze the calibration of the neural network. The calibration evaluations in terms of reliability diagrams and ECE are used to determine if the neural network's output probabilities represent the actual likelihood of the prediction being correct. In the original paper on neural network calibration [8], the authors claim that most modern CNNs are not well calibrated and produce overconfident predictions. In this review, we saw that several methods including TS and TTA can be applied to improve calibration [31].

Another observation we made is that combining uncertainty estimation methods can improve the results. This holds for combinations of Ensembles and MCDO [2,20,21], TS and MCDO [27], or TTA and MCDO [30].

By presenting retained data versus accuracy curves, several works [2,10,12,14,20,26,30] show that discarding uncertain predictions leads to an improved accuracy of the neural network on the remaining samples. This insight holds for all 3 categories of uncertainty estimation methods that we denoted as (1) model sampling, (2) single network methods, and (3) data augmentation. An important message from this observation is that uncertainty estimation can be used as a tool to improve the collaboration between AI systems and human experts. Thus far, all studies were performed in very artificial settings. Future work should therefore analyze the performance improvement of a collaboration between an uncertainty-aware AI system and human experts in scenarios that are closer to real-life situations in clinics.

Acknowledgments

The research is funded by the Ministerium für Soziales und Integration Baden Württemberg, Germany.

Authors' Contributions

AK, AH, and TJB are responsible for concept and design. AK and KH did the study selection. HM, EKH, JNK, SF, and CvK critically revised the manuscript and provided valuable feedback.

Conflicts of Interest

TJB is the owner of Smart Health Heidelberg GmbH (Handschuhsheimer Landstr. 9/1, 69120 Heidelberg, Germany, https://smarthealth.de) which develops telemedicine mobile apps (such as AppDoc; https://online-hautarzt.net and Intimarzt; https://intimarzt.de), outside of the submitted work.

References

1. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019 Jan 7;1(1):20-23. [doi: 10.1038/s42256-018-0004-1]
2. Filos A, Farquhar S, Gomez A, Rudner T, Kenton Z, Smith L, et al. A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. 2019 Presented at: Conference on Neural Information Processing Systems (NeurIPS); Dec 8-14; Vancouver, Canada URL: https://arxiv.org/pdf/1912.10481.pdf
3. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. 2019 Presented at: Annual Conference on Neural Information Processing Systems (NeurIPS); Dec 8-14; Vancouver, Canada URL: https://arxiv.org/pdf/1906.02530.pdf
4. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. 2016 Presented at: International Conference on Machine Learning (ICML); June 19-24; New York URL: https://arxiv.org/pdf/1506.02142.pdf
5. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. 2017 Presented at: Annual Conference on Neural Information Processing Systems (NeurIPS); Dec 4-9; Long Beach, CA URL: https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf
6. Rasmussen CE. Gaussian processes in machine learning. In: Advanced Lectures on Machine Learning. Heidelberg: Springer Berlin; 2003:63-71.
7. Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. In: Proceedings of the 32nd International Conference on Machine Learning. 2015 Presented at: PMLR; July 7-9; Lille, France URL: http://proceedings.mlr.press/v37/blundell15.pdf
8. Guo C, Pleiss G, Sun Y, Weinberger K. On calibration of modern neural networks. arXiv. Preprint posted online June 14, 2017 [FREE Full text] [doi: 10.48550/arXiv.1706.04599]
9. Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. 2018 Presented at: Annual Conference on Neural Information Processing Systems (NeurIPS); Dec 2-8; Montreal, Canada URL: https://arxiv.org/pdf/1806.01768.pdf
10. Ayhan M, Berens P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. 2018 Presented at: MIDL 2018; July 4-6; Amsterdam, the Netherlands URL: https://openreview.net/pdf?id=rJZz-knjz
11. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021 Mar;18(3):e1003583 [FREE Full text] [doi: 10.1371/journal.pmed.1003583] [Medline: 33780438]
12. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep 2017 Dec 19;7(1):17816 [FREE Full text] [doi: 10.1038/s41598-017-17876-z] [Medline: 29259224]
13. Laves M, Ihler S, Ortmaier T. Uncertainty quantification in computer-aided diagnosis: make your model say "I don't know" for ambiguous cases. 2019 Presented at: Conference on Medical Imaging with Deep Learning (MIDL); July 8-10; London, UK URL: https://openreview.net/pdf?id=rJevPsX854
14. Mobiny A, Singh A, Van Nguyen H. Risk-aware machine learning classifier for skin lesion diagnosis. J Clin Med 2019 Aug 17;8(8):1241 [FREE Full text] [doi: 10.3390/jcm8081241] [Medline: 31426482]
15. Herzog L, Murina E, Dürr O, Wegener S, Sick B. Integrating uncertainty in deep neural networks for MRI based stroke analysis. Med Image Anal 2020 Oct;65:101790. [doi: 10.1016/j.media.2020.101790] [Medline: 32801096]
16. Calderón-Ramírez S, Murillo-Hernández D, Rojas-Salazar K, Molina-Cabello M. Improving uncertainty estimations for mammogram classification using semi-supervised learning. 2021 Presented at: International Joint Conference on Neural Networks (IJCNN); July 18-22; Shenzhen, China. [doi: 10.1109/ijcnn52387.2021.9533719]
17. Calderon-Ramirez S, Yang S, Moemeni A, Colreavy-Donnelly S, Elizondo DA, Oala L, et al. Improving uncertainty estimation with semi-supervised deep learning for COVID-19 detection using chest X-ray images. IEEE Access 2021;9:85442-85454. [doi: 10.1109/access.2021.3085418]
18. Linmans J, van DLJ, Litjens G. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. 2020 Presented at: Medical Imaging with Deep Learning (MIDL); July 6-9; Montreal, Canada URL: https://geertlitjens.nl/publication/linm-20/linm-20.pdf
19. Thagaard J, Hauberg S, van DVB, Ebstrup T, Hansen J, Dahl A. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: Medical Image Computing and Computer Assisted Intervention. 2020 Presented at: MICCAI; Oct 4-8; Lima, Peru p. 824-833. [doi: 10.1007/978-3-030-59710-8_80]

20. Yang S, Fevens T. Uncertainty quantification and estimation in medical image classification. In: Artificial Neural Networks and Machine Learning. 2021 Presented at: ICANN 2021; Sep 14-17; Bratislava, Slovakia p. 671-683. [doi: 10.1007/978-3-030-86365-4_54]
21. Abdar M, Samami M, Dehghani Mahmoodabad S, Doan T, Mazoure B, Hashemifesharaki R, et al. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. Comput Biol Med 2021 Aug;135:104418. [doi: 10.1016/j.compbiomed.2021.104418] [Medline: 34052016]
22. Berger C, Paschali M, Glocker B, Kamnitsas K. Confidence-based out-of-distribution detection: a comparative study and analysis. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. 2021 Presented at: UNSURE 2021; Oct 1; Strasbourg, Farnce p. 122-132. [doi: 10.1007/978-3-030-87735-4_12]
23. Toledo-Cortés S, de LPM, Perdomo O, González F. Hybrid deep learning Gaussian process for diabetic retinopathy diagnosis and uncertainty quantification. In: Ophthalmic Medical Image Analysis. 2020 Presented at: OMIA 2020; Oct 8; Lima, Peru p. 206-215. [doi: 10.1007/978-3-030-63419-3_21]
24. Ghesu F, Georgescu B, Gibson E, Guendel S, Kalra M, Singh R, et al. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In: Medical Image Computing and Computer Assisted Intervention. 2019 Presented at: MICCAI; Oct 13-17; Shenzhen, China p. 676-684. [doi: 10.1007/978-3-030-32226-7_75]
25. Tardy M, Scheffer B, Mateus D. Uncertainty measurements for the reliable classification of mammograms. In: Medical Image Computing and Computer Assisted Intervention. 2019 Presented at: MICCAI; Oct 13-17; Shenzhen, China p. 495-503. [doi: 10.1007/978-3-030-32226-7_55]
26. Ghesu FC, Georgescu B, Mansoor A, Yoo Y, Gibson E, Vishwanath RS, et al. Quantifying and leveraging predictive uncertainty for medical image assessment. Med Image Anal 2021 Feb;68:101855. [doi: 10.1016/j.media.2020.101855] [Medline: 33260116]
27. Carneiro G, Zorron Cheng Tao Pu L, Singh R, Burt A. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Med Image Anal 2020 May;62:101653. [doi: 10.1016/j.media.2020.101653] [Medline: 32172037]
28. Liang G, Zhang Y, Jacobs N. Neural network calibration for medical imaging classification using DCA regularization. 2020 Presented at: International Conference on Machine Learning (ICML); July 17; Virtual workshop URL: http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-137.pdf
29. Jensen M, Jørgensen D, Jalaboi R, Hansen M, Olsen M. Improving uncertainty estimation in convolutional neural networks using Inter-rater agreement. In: Medical Image Computing and Computer Assisted Intervention. 2019 Presented at: MICCAI; Oct 13-17; Shenzhen, China p. 540-548. [doi: 10.1007/978-3-030-32251-9_59]
30. Combalia M, Hueto F, Puig S, Malvehy J, Vilaplana V. Uncertainty estimation in deep neural networks for dermoscopic image classification. 2020 Presented at: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 14-19; Seattle, WA. [doi: 10.1109/cvprw50498.2020.00380]
31. Ayhan MS, Kühlewein L, Aliyeva G, Inhoffen W, Ziemssen F, Berens P. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. Med Image Anal 2020 Aug;64:101724. [doi: 10.1016/j.media.2020.101724] [Medline: 32497870]
32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. Preprint posted online Sep 4, 2014 [FREE Full text] [doi: 10.48550/arXiv.1907.11692]
33. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30; Las Vegas, NA. [doi: 10.1109/CVPR.2016.90]
34. Huang G, Liu Z, van DML, Weinberger K. Densely connected convolutional networks. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26; Honolulu, HA. [doi: 10.1109/cvpr.2017.243]
35. Liang S, Li Y, Srikant R. Enhancing the reliability of out-of-distribution image detection in neural networks. 2018 Presented at: 6th International Conference on Learning Representations (ICLR); Apr 30 - May 3; Vancouver, Canada URL: https://openreview.net/pdf?id=H1VGkIxRZ
36. Szegedy C, Vanhoucke V, Ioffe S. Rethinking the inception architecture for computer vision. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30; Las Vegas, NA. [doi: 10.1109/cvpr.2016.308]
37. Ding Y, Liu J, Xiong J, Shi Y. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. 2020 Presented at: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 14-19; Seattle, WA. [doi: 10.1109/cvprw50498.2020.00010]

Abbreviations

AI: artificial intelligence
AUPRC: area under the precision recall curve
AUROC: area under the receiver operating curve
CNN: convolutional neural network
DCA: difference between confidence and accuracy

ECE: expected calibration error
EDL: evidential deep learning
GP: Gaussian process
MCDO: Monte Carlo dropout
MFVI: mean field variational inference
MRI: magnetic resonance imaging
OOD: out-of-distribution
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SVI: stochastic variational inference
TS: temperature scaling
TTA: test-time data augmentation

Edited by C Lovis; submitted 14.01.22; peer-reviewed by E Rezk, G Nneji; comments to author 08.03.22; revised version received 11.04.22; accepted 04.06.22; published 02.08.22.

Please cite as:
Kurz A, Hauser K, Mehrtens HA, Krieghoff-Henning E, Hekler A, Kather JN, Fröhling S, von Kalle C, Brinker TJ
Uncertainty Estimation in Medical Image Classification: Systematic Review
JMIR Med Inform 2022;10(8):e36427
URL: https://medinform.jmir.org/2022/8/e36427
doi:10.2196/36427
PMID:35916701

Review

State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review

Georgios Petmezas[1], BSc, MSc; Leandros Stefanopoulos[1], BSc, MSc; Vassilis Kilintzis[1], BSc, MSc, PhD; Andreas Tzavelis[2], BSc; John A Rogers[3], BSc, MSc, PhD; Aggelos K Katsaggelos[4], PhD; Nicos Maglaveras[1], MSc, PhD

[1]Lab of Computing, Medical Informatics and Biomedical-Imaging Technologies, The Medical School, Aristotle University of Thessaloniki, Thessaloniki, Greece

[2]Department of Biomedical Engineering, Northwestern University, Evanston, IL, United States

[3]Department of Material Science, Northwestern University, Evanston, IL, United States

[4]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, United States

Corresponding Author:
Nicos Maglaveras, MSc, PhD
Lab of Computing, Medical Informatics and Biomedical-Imaging Technologies
The Medical School
Aristotle University of Thessaloniki
University Campus - Box 323
Thessaloniki, 54124
Greece
Phone: 30 2310999281
Email: nicmag@auth.gr

Abstract

Background: Electrocardiogram (ECG) is one of the most common noninvasive diagnostic tools that can provide useful information regarding a patient's health status. Deep learning (DL) is an area of intense exploration that leads the way in most attempts to create powerful diagnostic models based on physiological signals.

Objective: This study aimed to provide a systematic review of DL methods applied to ECG data for various clinical applications.

Methods: The PubMed search engine was systematically searched by combining "deep learning" and keywords such as "ecg," "ekg," "electrocardiogram," "electrocardiography," and "electrocardiology." Irrelevant articles were excluded from the study after screening titles and abstracts, and the remaining articles were further reviewed. The reasons for article exclusion were manuscripts written in any language other than English, absence of ECG data or DL methods involved in the study, and absence of a quantitative evaluation of the proposed approaches.

Results: We identified 230 relevant articles published between January 2020 and December 2021 and grouped them into 6 distinct medical applications, namely, blood pressure estimation, cardiovascular disease diagnosis, ECG analysis, biometric recognition, sleep analysis, and other clinical analyses. We provide a complete account of the state-of-the-art DL strategies per the field of application, as well as major ECG data sources. We also present open research problems, such as the lack of attempts to address the issue of blood pressure variability in training data sets, and point out potential gaps in the design and implementation of DL models.

Conclusions: We expect that this review will provide insights into state-of-the-art DL methods applied to ECG data and point to future directions for research on DL to create robust models that can assist medical experts in clinical decision-making.

(JMIR Med Inform 2022;10(8):e38454) doi:10.2196/38454

KEYWORDS

electrocardiogram; ECG; ECG databases; deep learning; convolutional neural networks; CNN; residual neural network; ResNet; long short-term memory; LSTM; diagnostic tools; decision support; clinical decision

Introduction

Study Background

Electrocardiogram (ECG) is one of the most common noninvasive diagnostic tools used in clinical medicine [1]. An ECG is a nonstationary physiological signal that measures voltage changes produced by the electrical activity of the heart. It is mostly used by cardiologists to assess heart function and electrophysiology [2]. ECG interpretation plays a vital role in personalized medicine and can assist in cardiovascular disease (CVD) detection, rehabilitation, and the development of treatment strategies. Owing to the major increase in the amount of ECG data available and measurement heterogeneity from medical devices and placements, there are many cases where traditional diagnosis becomes inefficient, as it requires complex manual analysis and highly trained medical experts to achieve adequate accuracy [3].

During the past few decades, the massive surge in computational power and availability of large data sets have created new opportunities for machine-driven diagnosis in many health care areas [4]. Artificial intelligence (AI) is leading the way in most attempts to develop reliable diagnostic tools based on data-driven techniques [5]. In particular, deep learning (DL) algorithms, a subset of machine learning (ML), can generate powerful models that can learn relationships between data and reveal hidden patterns in complex biomedical data without the need for prior knowledge. DL models adjust better to large data sets and, in most cases, continue to improve with the addition of more data, thus enabling them to outperform most classical ML approaches [6,7]. They have been tested extensively in many application areas, such as speech recognition, visual object recognition, object detection, and natural language processing, achieving promising results [8].

DL algorithms are typically based on deep network architectures comprising multiple hidden layers [9]. The most frequently used DL algorithms are convolutional neural networks (CNNs), which were originally proposed for object recognition and image classification [10,11]. Since then, they have been successfully used in various medical applications, including medical image analysis [12], biomedical signal classification [13,14], pulmonary sound classification [15], biomedical signal quality assessment [16], pathological voice detection [17], and sleep staging [18].

Moreover, residual neural networks (ResNets) [19], which were recently proposed to solve the difficulties of training very deep neural networks (DNNs), are well established and used in several medical tasks, such as prostate cancer detection [20], nuclei segmentation and detection [21], coronary calcium detection [22], and pulmonary nodule classification [23].

In addition to CNN and ResNet architectures, recurrent neural networks (RNNs) represent another type of DL technique frequently used in health care. Disease prediction [24], biomedical image segmentation [25], and obstructive sleep apnea detection [26] are only a few of their applications. More specifically, the performance of improved versions of classic RNNs, such as long short-term memory (LSTM) networks and gated recurrent units (GRUs), has been studied extensively in recent years in a series of health-related tasks, including medical image denoising [27], Alzheimer disease detection [28], life expectancy prediction [29], cardiac arrhythmia classification [30], epileptic seizure detection [31], cell segmentation [32], and cardiac phase detection [33].

Another DL method proposed in 2017 that has recently gained popularity among the scientific community is transformers [34], which adopts the mechanism of self-attention to handle sequential data. They have been tested in a series of medical tasks, including cardiac abnormality diagnosis [35], food allergen identification [36], medical language understanding [37], and chemical image recognition [38].

Finally, autoencoders, a DNN technique capable of learning compressed representations of its inputs, have been tested in several medical applications, such as the prediction of heart transplant rejection [39], cell detection and classification [40], anticancer drug response classification [41], premature ventricular contraction detection [42], and endomicroscopic image classification [43].

The purpose of this study is to provide a complete and systematic account of the current state-of-the-art DL methods for ECG data. The main idea behind this comprehensive review is to group and summarize the DL approaches per field of application, discuss the most notable studies, and provide a detailed overview of the major ECG databases. In addition, we will identify important open research problems and directions and provide an assessment of the future of the field. We expect this review to be of great value to newcomers to the topic, as well as to practitioners in the field.

The remainder of this paper is structured as follows: In the *Background of DL* section, background knowledge for DL techniques and algorithms is presented, and related state-of-the-art methods for ECG processing and analysis are reviewed. In the *Methods* section, the research methodology is described in detail, and, in the *Results* section, the results of the systematic review are presented. In the *Discussion* section, a discussion based on the research findings is presented. Finally, the conclusions of the study are summarized in the *Conclusions* section.

Background of DL

DL Algorithm

DL is a branch of ML that uses multilayered structures of algorithms called neural networks (NNs) to learn representations of data by using multiple levels of abstraction [8]. Unlike most traditional ML algorithms, many of which have a finite capacity to learn regardless of how much data they acquire, DL systems can usually improve their performance with access to more data.

Given the availability of large data sets and advancements in modern technology, DL has seen a spectacular rise in the past decade. DL algorithms can construct robust data-driven models that can reveal hidden patterns in data and make predictions based on them. The following subsections describe some of the most commonly used DL methods that are applied to a wide range of health-related tasks where ECG data are present.

CNN Algorithm

CNNs are among the most popular DL architectures and owe their name to the mathematical concept of convolution. CNNs are designed to adaptively learn the spatial hierarchy of data by extracting and memorizing high- and low-level patterns to predict the final output.

Although they were initially designed to deal with 2D image data [44], during the past few years, several modified 1D versions of them have been proposed for numerous applications, achieving state-of-the-art performance [45].

The structure of a typical CNN integrates a pipeline of multiple hidden layers, in particular, convolutional and pooling layers, followed by fully connected layers. The convolutional layers implement filters (or kernels) that perform convolution between the kernel (impulse response of the filter) and the input signal. In this way, each convolutional layer creates features (or activation maps) from its input, a process commonly known as feature extraction.

In contrast, the pooling layers conduct down-sampling of the extracted feature maps to reduce the computational complexity required when processing large volumes of data. Finally, the fully connected layers are simple feed-forward NNs that create weighted connections between successive layers. Therefore, they achieve the mapping of the aggregated activations of all previous layers into a class probability distribution by applying a sigmoid or *softmax* activation function that represents the final output of the CNN.

ResNet Algorithm

ResNet is a special type of DL network that was proposed to solve the vanishing gradient problem, which occurs when training DNNs. In other words, as the number of stacked layers of a DNN increases, the gradient of the earlier layers vanishes. Thus, the network fails to update the weights of the earlier layers. This means that no learning occurs in the earlier layers, resulting in poor training and testing performance.

The key idea behind ResNet is the introduction of residual blocks that use skip connections to add the outputs from earlier layers to those of later layers. Precisely, the network creates shortcuts that enable the gradient to take shorter paths through the deeper layers, thereby eliminating the vanishing gradient problem. Thus, the precision of deep feature extraction is improved, whereas the computational complexity of the network remains substantially low.

ResNet is typically a network comprising CNN blocks that are successively repeated multiple times. Many variants of the ResNet architecture use the same concept but various numbers of layers to address different problems, such as ResNet-34, ResNet-50, and ResNet-101, where 34, 50, and 101 are the depths of the network, respectively.

RNN Algorithm

RNNs were first introduced by Rumelhart et al [46] in 1986. They are a class of artificial NNs capable of memorizing the temporal dynamics of sequential data by forming a directed graph along them. Specifically, they deploy hidden units that create strong dependencies among data by preserving valuable information regarding previous inputs to predict current and future outputs.

However, as the time distance between dependent inputs increases, RNNs become incapable of handling long-term dependencies because of the vanishing gradient problem. To address this problem, new variations of RNNs have been proposed, including LSTM networks and GRUs.

LSTM networks were introduced by Hochreiter and Schmidhuber [47] in 1997. They solved the problem of long-term dependencies by implementing gates to control the memorization process. This means that they can recognize and retain both the long- and short-term dependencies between the data of a sequential input for long periods, resulting in efficient learning and, finally, improved performance.

The structure of LSTM comprises an ordered chain of identical cells. Each cell is responsible for transferring 2 states to the next cell, namely, the current internal cell state and its internal hidden state, also known as short-term and long-term memory, respectively. To achieve this, it uses 3 types of gates, namely forget, input, and output gates, to control the information that is passed onto further computations.

Specifically, using the forget gate, the cell determines which part of the previous time stamp's information needs to be retained and which should be forgotten. The input gate updates the cell state by adding new information. Finally, the output gate selects information that will be passed on as the output of the cell. By controlling the process of adding valuable information or removing unnecessary information, a cell can remember long-term dependencies over arbitrary time intervals.

In contrast, motivated by the LSTM unit, in 2014, Cho et al [48] proposed GRUs to address the vanishing gradient problem. Unlike LSTMs, GRUs do not have separate cell states. In addition, they use only 2 gates to control the flow of information via the hidden state, namely, the update and reset gates.

Precisely, the update gate, which acts as the unit's long-term memory, is responsible for selecting the amount of previous information that must be passed on to the current hidden state. By contrast, the reset gate represents the short-term memory of the unit and oversees the determination of the amount of past information that must be ignored.

With these 2 gates, each hidden unit can capture dependencies over different time scales. Thus, units trained to capture long-term dependencies tend to have update gates that are mostly active, and conversely, those trained to memorize short-term dependencies tend to have active reset gates.

Autoencoders

Autoencoders are a special type of feed-forward NNs that was introduced by Rumelhart et al [49] in 1986. An autoencoder can learn efficient representations of data and is mainly applied for feature extraction and dimensionality reduction.

A typical autoencoder structure includes 2 parts: encoder and decoder. The encoder compresses the input and creates a latent representation, which is mapped to a hidden layer, also known

as a bottleneck. Then, the decoder uses this latent representation to reconstruct the original input.

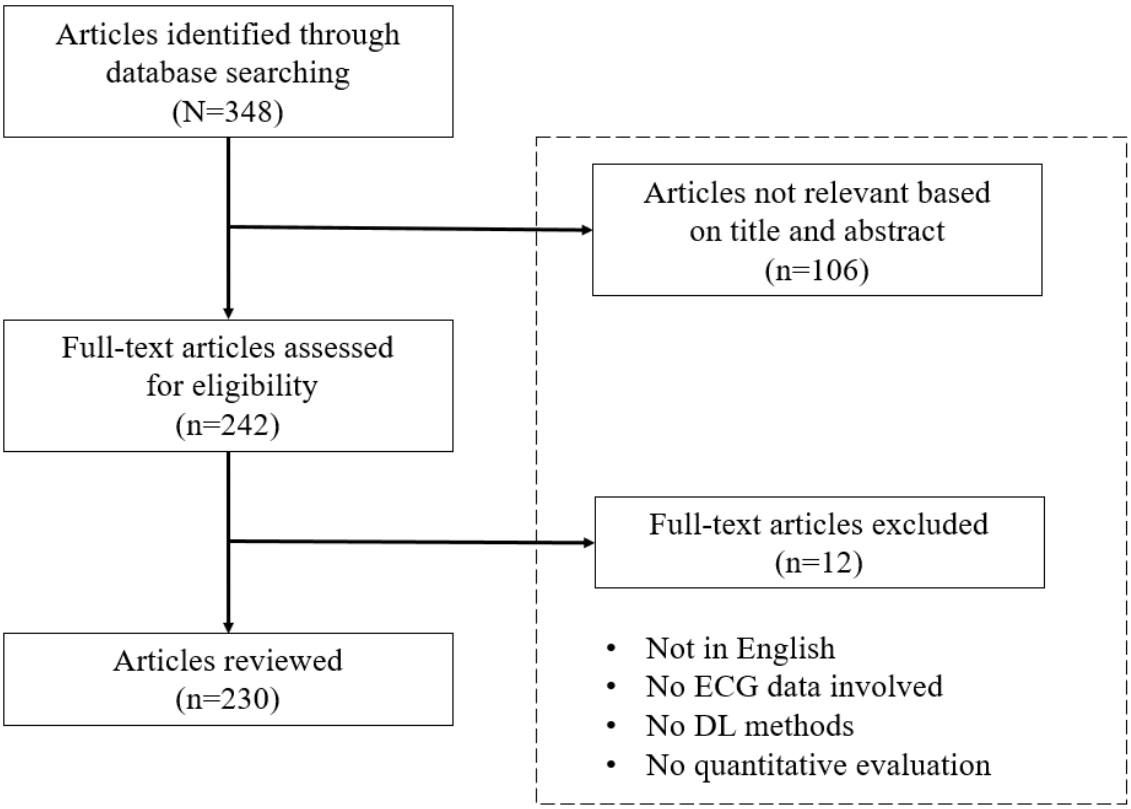
In this manner, an autoencoder is trained by minimizing the reconstruction error to learn to create low-dimensional copies of higher-dimensional data. There are several types of autoencoders, including denoising autoencoders [50], variational autoencoders [51], and convolutional autoencoders [52].

Methods

Literature Search

The PubMed search engine was systematically searched by combining "deep learning" and keywords such as "ecg," "ekg," "electrocardiogram," "electrocardiography," and "electrocardiology." During the initial screening, 348 unique articles published in various journals between January 2020 and December 2021 were identified. Of these 348 articles, 106 (30.5%) were excluded based on their titles and abstracts, and the remaining 242 (69.5%) were further reviewed. The reasons for article exclusion were manuscript written in any language other than English, absence of ECG data or DL methods involved in the study, and absence of a quantitative evaluation of the proposed approaches. After a full-text assessment, 4.9% (12/242) of the articles were excluded as they were about works that did not include ECG signals. Finally, 230 relevant articles were selected for this review. The detailed process of the literature search and selection is illustrated in Figure 1.

Figure 1. Flow diagram of the literature search. DL: deep learning; ECG: electrocardiogram.



Bibliometric Analysis

To obtain a clear picture of the literature search results, a co-occurrence analysis was conducted. For this purpose, the VOSviewer software tool (Nees Jan van Eck and Ludo Waltman) [53] was used to create and visualize 3 maps based on the bibliographic data of this study. Specifically, all keywords from the 230 relevant studies were grouped and linked to establish the impact of each keyword on the given scientific field and its interconnections with other keywords. In this way, 3 distinct clusters of keywords were formed, namely "clinical issues" (cluster 1), "methods and tools" (cluster 2), and "study characteristics" (cluster 3), as shown in Textbox 1, and an individual map was generated for each of the 3 categories. Figure 2 displays the co-occurrence network that corresponds to the "clinical issues" cluster of keywords. Cardiac arrhythmias and atrial fibrillation (AF) were identified as the major clinical issues in this review. Figure 3 presents the co-occurrence network for the "methods and tools" cluster, where ECG and DL constitute the network's core. Finally, Figure 4 shows the co-occurrence network for the "study characteristics" cluster, where, as expected, humans are the center of attention.

Textbox 1. Keyword cluster summary.

Cluster and keywords

- Cluster 1
  - "arrhythmias, cardiac," "atrial fibrillation," "biometric identification," "blood pressure determination," "cardiomyopathy," "cardiovascular diseases," "coronary artery disease," "covid-19," "early diagnosis," "fetal monitoring," "heart diseases," "heart failure," "heartbeat classification," "hypertension," "monitoring, physiologic," "myocardial infarction," "sleep apnea," "sudden cardiac death," "ventricular fibrillation," "ventricular function, left," "ventricular premature complexes"
- Cluster 2
  - "12-lead ecg," "algorithms," "artificial intelligence," "attention mechanism," "blood pressure," "cardiology," "continuous wavelet transform," "convolutional neural networks, computer," "data compression," "deep learning," "deep neural networks, computer," "diagnosis, computer-assisted," "echocardiography," "electrocardiography," "electroencephalography," "feature extraction," "feature fusion," "heart," "heart rate," "heart rate variability," "long short-term memory," "machine learning," "neural networks, computer," "photoplethysmography," "polysomnography," "recurrent neural networks, computer," "signal processing, computer-assisted," "supervised machine learning," "support vector machine," "wavelet analysis," "wearable electronic devices"
- Cluster 3
  - "adult," "aged," "aged, 80 and over," "cohort studies," "databases, factual," "female," "humans," "male," "middle aged," "predictive value of tests," "pregnancy," "reproducibility of results," "retrospective studies," "roc curve," "sensitivity and specificity," "young adult"

Figure 2. The co-occurrence network for the "clinical issues" cluster.

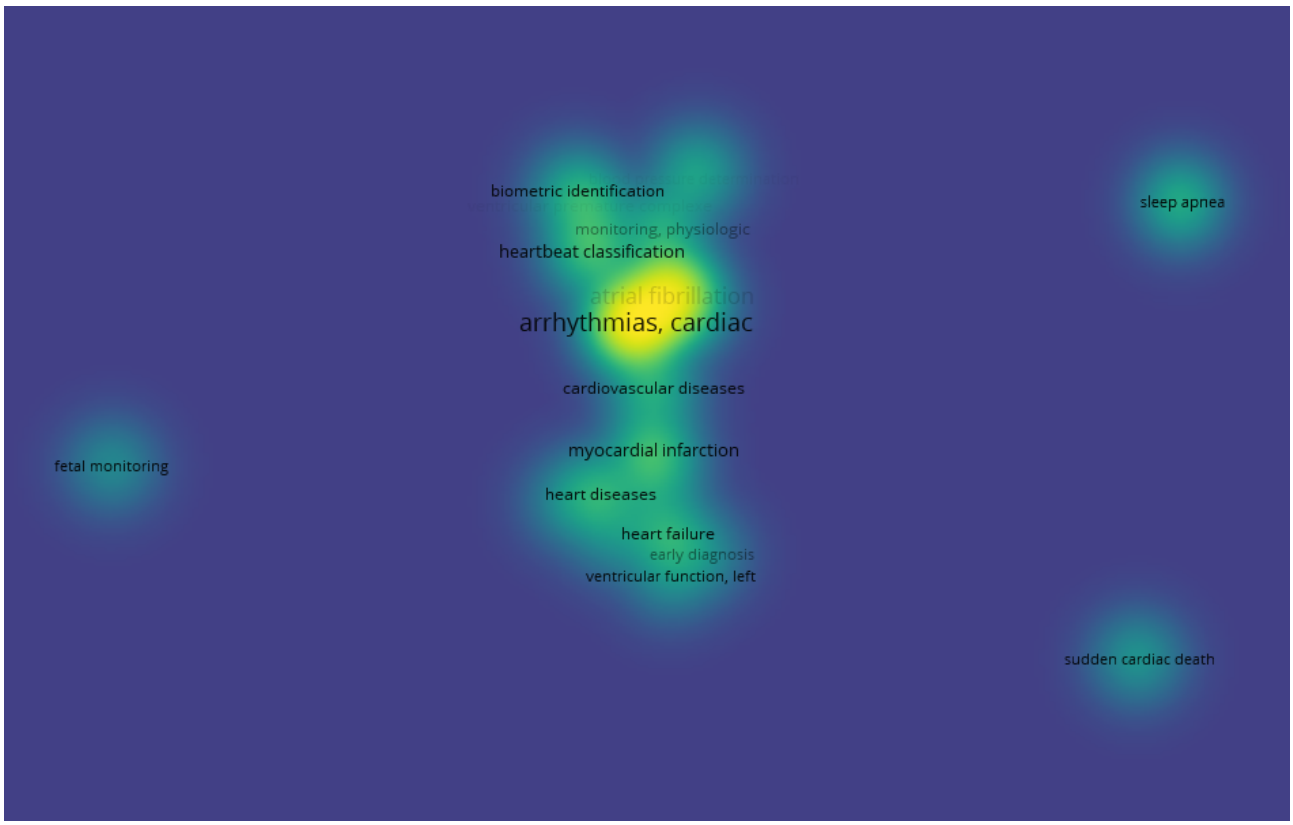


Figure 3. The co-occurrence network for the “methods and tools” cluster.

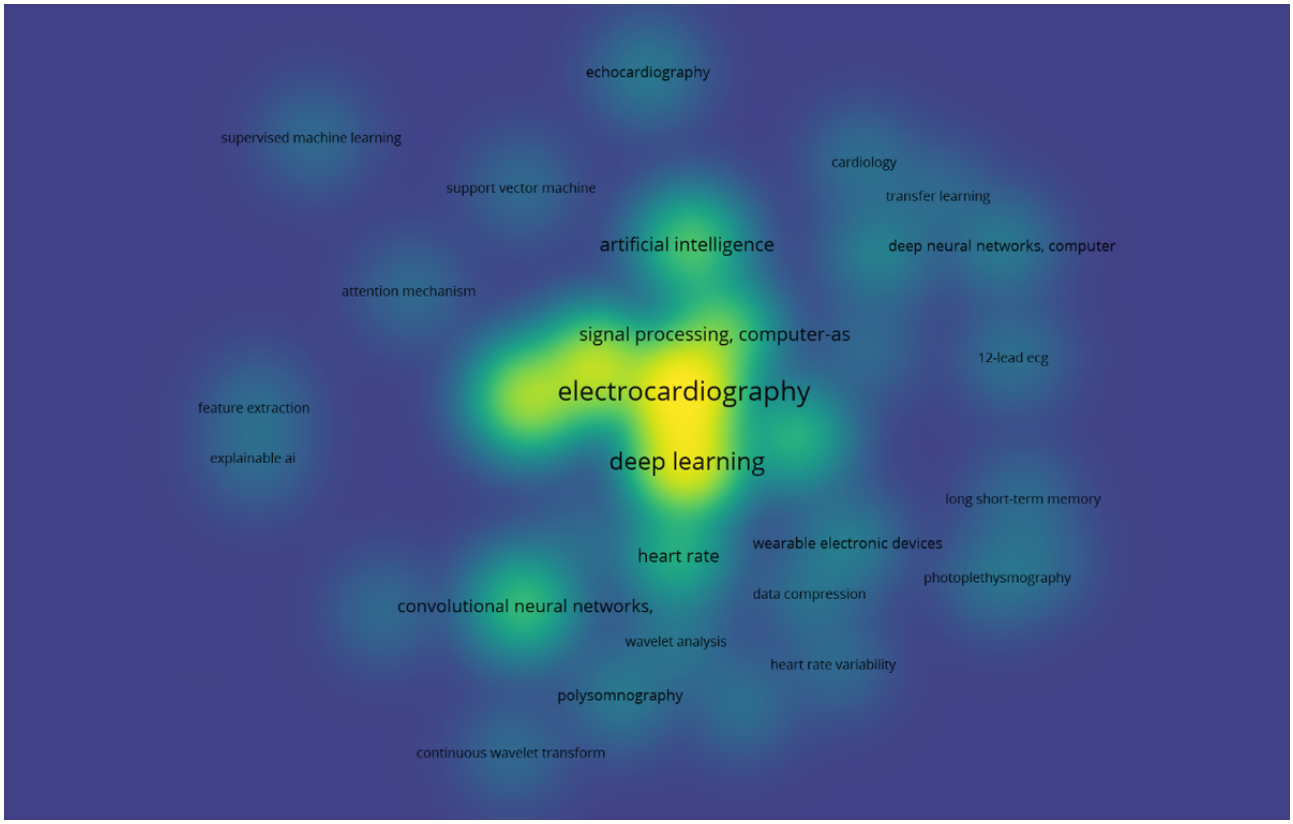
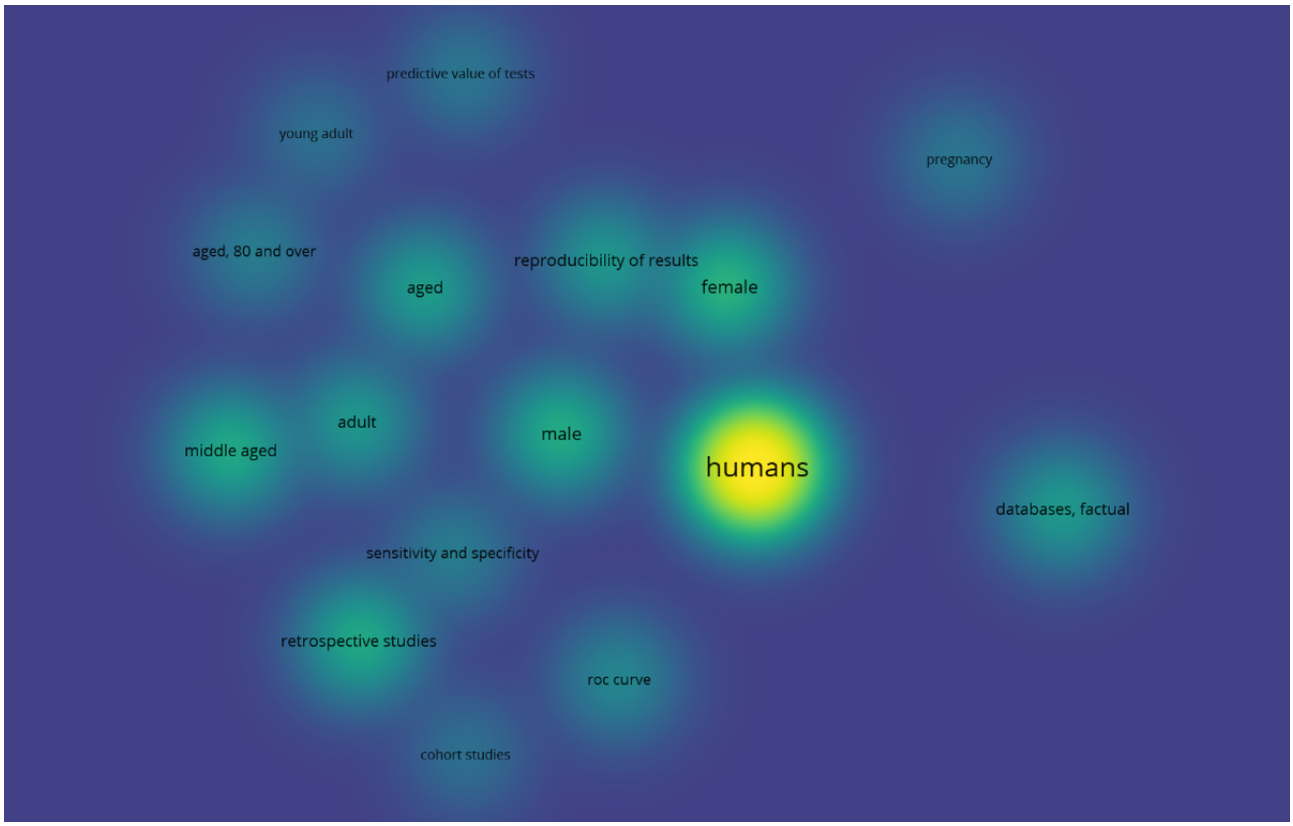


Figure 4. The co-occurrence network for the “study characteristics” cluster.



Results

ECG Data Sources

On the basis of the selected studies, multiple ECG data sources were identified, including several well-established publicly available databases. These data sources exhibit differences in the number of enrolled patients, number of recordings, ECG systems used to collect them, data duration, and sample rate. Their content is presented in Multimedia Appendix 1 [54-92], where the links to publicly available data are placed as hyperlinks on the name of each database.

The most commonly used databases were the Massachusetts Institute of Technology (MIT)–Beth Israel Hospital (BIH) Arrhythmia Database [80] (55/230, 23.9% studies), 2017 PhysioNet/CinC Challenge database [57] (31/230, 13.5% studies), the China Physiological Signal Challenge (CPSC) 2018 database [69] (26/230, 11.3% studies), the MIT-BIH Atrial Fibrillation Database [81] (17/230, 7.4% studies), and the *Physikalisch Technische Bundesanstalt* (PTB)–XL ECG data set [87] (17/230, 7.4% studies).

The MIT-BIH Arrhythmia Database contains 48 half-hour excerpts of 2-channel ambulatory ECG recordings obtained from 47 participants studied by the BIH Arrhythmia Laboratory between 1975 and 1979 with a sampling frequency of 360 Hz. Of these, 23 recordings were chosen at random from a set of 4000 recordings of 24-hour ambulatory ECG collected from a mixed population of inpatients (approximately 60%) and outpatients (approximately 40%) at Boston's BIH, whereas the remaining 25 recordings were selected from the same set to include less common but clinically significant arrhythmias that would not be well represented in a small random sample. Finally, each recording was independently annotated by ≥2 cardiologists.

In contrast, the 2017 PhysioNet/CinC Challenge database contains 12,186 single-lead ECG recordings collected using a sampling frequency of 300 Hz. The training set contains 8528 single-lead ECG recordings lasting from 9 seconds to just >60 seconds, and the test set contains 3658 ECG recordings of similar lengths.

The CPSC 2018 database comprises ECG recordings collected from 11 hospitals by using a sampling frequency of 500 Hz. The training set contains 6877 (female: 3178; male: 3699) 12-lead ECG recordings lasting from 6 seconds to 60 seconds, and the test set, which is unavailable to the public for scoring purposes, contains 2954 ECG recordings of similar lengths.

Furthermore, the MIT-BIH Atrial Fibrillation Database includes 25 long-term ECG recordings of human patients with AF (mostly paroxysmal). The individual recordings are each 10 hours in duration and contain 2 ECG signals, each sampled at 250 Hz, whereas the rhythm annotation files were manually prepared and contain rhythm annotations of 4 types, namely, AFIB (AF), AFL (atrial flutter), J (AV junctional rhythm), and N (all other rhythms).

Finally, the PTB-XL ECG data set is a large data set of 21,837 clinical 12-lead ECGs from 18,885 patients with a duration of 10 seconds and a sampling frequency of 500 Hz. The raw waveform data were annotated by up to 2 cardiologists who assigned multiple ECG statements to each record.

Medical Applications

Overview

The 230 relevant articles identified during the literature search were grouped into several categories based on their study objectives. In particular, 6 distinct medical applications were identified: blood pressure (BP) estimation, CVD diagnosis, ECG analysis, biometric recognition, sleep analysis, and other clinical analyses.

Most of the studies use ECG signals for CVD diagnosis, mainly via signal or beat classification. Moreover, a significant portion of them uses DL algorithms to perform ECG analysis, as well as diagnosis of other clinical conditions.

In this study, the identified DL approaches are grouped per field of application, and the most notable approaches are discussed in detail. Moreover, Multimedia Appendix 2 [93-322] provides details regarding the author and the year of publication of each article, the medical task that each article refers to, data, data preprocessing, splitting strategy, DL algorithm applied in each study, and performance of each approach.

BP Estimation

Only 2.6% (6/230) of studies that applied DL methods to ECG data to perform BP estimation were identified in the literature search. A combined architecture of ResNets and LSTM was proposed twice (33.3%), once by Miao et al [94], who achieved a mean error of −0.22 (SD 5.82) mm Hg for systolic BP (SBP) prediction and of −0.75 (SD 5.62) mm Hg for diastolic BP (DBP) prediction using data that originated from a private database, and once by Paviglianiti et al [96], who achieved a mean average error of 4.118 mm Hg for SBP and 2.228 mm Hg for DBP prediction using the Medical Information Mart for Intensive Care database. By contrast, Jeong and Lim [98] exercised a CNN-LSTM network on the Medical Information Mart for Intensive Care database and managed to predict SBP and DBP with a mean error of 0.0 (SD 1.6) mm Hg and 0.2 (SD 1.3) mm Hg, respectively.

CVD Diagnosis

More than half (152/230, 66.1%) of the studies that were identified during the literature search applied DL methods to ECG data for CVD diagnosis. The most common data sources for CVD diagnosis are private (37%) and mixed public (25%) databases. However, a notable proportion (15%) of the aforementioned studies exclusively used the MIT-BIH Arrhythmia Database. Almost the half of them (10/23, 43.5%) applied a CNN structure.

Regarding the MIT-BIH Arrhythmia Database, the best accuracy (99.94%) was achieved by Wang et al [185], who introduced a fused autoencoder-CNN network to classify 6 different ECG rhythms. However, a high percentage of the studies that managed to classify data originating from the same database implemented a CNN structure. Lu et al [180] used a 1D-CNN for arrhythmia classification, achieving an accuracy of 99.31%,

whereas Yu et al [219] used a 1D-CNN to detect premature ventricular contraction, achieving a classification accuracy of 99.70%.

On the contrary, a ResNet architecture was tested only 3 times on the MIT-BIH Arrhythmia Database; nonetheless, 0.9% (2/230) of these studies showed a high model performance. In particular, Li et al [146] proposed a ResNet model for heartbeat classification, achieving a classification accuracy of 99.38%, whereas Zhang et al [211] used a ResNet-101 structure to classify ECG beats with transfer learning and achieved an accuracy of 99.75%.

Regarding the rest of the databases, several noteworthy studies were identified in the literature. Specifically, Cai et al [101] implemented a densely connected DNN on a private ECG database for AF detection, achieving an accuracy between 97.74% and 99.35% for 3 different classification tasks, whereas Ghosh et al [103] applied a hierarchical extreme learning machine to ECG data from multiple public databases, achieving an accuracy of 99.40% in detecting AF.

Furthermore, Butun et al [125] proposed a 1D-capsule NN for the detection of coronary artery disease, achieving classification accuracies of 99.44% and 98.62% on 2-second and 5-second ECG segments, respectively, originating from a private ECG database. Another study by Thiagarajan et al [129] used multiple convolutional and pooling layers within a structure named DDxNet on ECG data from 2 public databases, achieving an accuracy of 98.50% for arrhythmia classification and 99.90% for myocardial infarction detection.

A study by Radhakrishnan et al [163] evaluated the performance (sensitivity 99.17%, specificity 99.18%, and accuracy 99.18%) of a 2D bidirectional LSTM network to detect AF in ECG signals from 4 public databases, whereas Petmezas et al [170] tested (sensitivity 97.87% and specificity 99.29%) a CNN-LSTM model on ECG signals originating from the MIT-BIH Atrial Fibrillation Database for the same medical task.

Moreover, Jahmunah et al [192] applied a CNN architecture to ECG data from several public ECG databases to detect coronary artery disease, myocardial infarction, and congestive heart failure, achieving an accuracy of 99.55%. Another study by Dai et al [195] proposed a CNN for CVD diagnosis using different intervals of ECG signals from the PTB Diagnostic ECG Database and achieved accuracies of 99.59%, 99.80%, and 99.84% for 1-, 2-, and 3-second ECG segments, respectively.

Finally, Ma et al [208] introduced an improved dilated causal CNN to classify ECG signals from the MIT-BIH Atrial Fibrillation Database, achieving a high model performance (sensitivity 98.79%, specificity 99.04%, and accuracy 98.65%), whereas Zhang et al [238] tested (sensitivity 99.65%, specificity 99.98%, and accuracy 99.84%) a CNN for AF detection on ECG signals from 2 major public databases.

ECG Analysis

In total, 12.6% (29/230) of studies that applied DL methods to ECG data to perform ECG analysis were identified during the literature search. Once again, CNN was the most commonly used DL method (11/29, 38%); nonetheless, the best model accuracy was achieved by studies using other DL methods. In particular, Teplitzky et al [251] tested (sensitivity 99.84% and positive predictive value 99.78%) a hybrid approach that combines 2 DL approaches, namely BeatNet and RhythmNet, to annotate ECG signals that originated from both public and private ECG databases, whereas Murat et al [258] used a CNN-LSTM approach on ECG data from the MIT-BIH Arrhythmia Database and achieved an accuracy of 99.26% in detecting 5 types of ECG beats.

By contrast, Vijayarangan et al [261] used a fused CNN-ResNet structure to perform R peak detection in ECG signals from several public ECG databases and achieved F_1-scores between 96.32% and 99.65% for 3 testing data sets. Another study by Jimenez Perez et al [265] implemented a U-Net model to delineate 2-lead ECG signals originating from the QT Database and achieved sensitivities of 98.73%, 99.94%, and 99.88% for P wave, QRS complex, and T wave detection, respectively. Finally, a study by Strodthoff et al [274] used a ResNet for patient sex identification by using 12-lead ECG recordings lasting between 6 and 60 seconds from several public databases and achieved an area under the curve of 0.925 for the PTB-XL ECG data set and 0.974 for the CPSC 2018 database.

Biometric Recognition

Only 3% (7/230) of studies that applied DL methods to ECG data to perform biometric recognition were identified in the literature search. Although 57% (4/7) of the studies used a CNN architecture, only 29% (2/7) of them achieved high model performance. Specifically, Wu et al [284] achieved an identification rate of >99% by using ECG signals from 2 public databases, whereas Chiu et al [285] achieved an identification rate of 99.10% by using single-lead ECG recordings that originated from the PTB Diagnostic ECG Database.

On the contrary, Song et al [281] implemented a ResNet-50 architecture for person identification using multiple ECG, face, and fingerprint data from several public and private databases and achieved an accuracy of 98.97% for ID classification and 96.55% for gender classification. Finally, AlDuwaile and Islam [283] tested several pretrained models, including GoogleNet, ResNet, MobileNet, and EfficientNet, and a CNN model to perform human recognition using ECG signals that originated from 2 public databases and achieved an accuracy between 94.18% and 98.20% for ECG-ID mixed-session and multisession data sets.

Sleep Analysis

Approximately 5.2% (12/230) of studies that applied DL methods to ECG data to perform sleep analysis were identified during the literature search. Half (6/12, 50%) of the studies proposed a CNN model, some of which achieved high performance in several sleep analysis–related tasks. In particular, Chang et al [289] used 1-minute ECG segments from the Apnea-ECG Database and designed a CNN to detect sleep apnea, achieving an accuracy of 87.90% and 97.10% for per-minute and per-recording classification, respectively.

In addition, a study by Urtnasan et al [291] proposed a CNN for the identification of sleep apnea severity by using ECG

segments from a private database and achieved an F_1-score of 98.00%, whereas another study by Urtnasan et al [297] implemented a CNN to classify sleep disorders by using polysomnography recordings from the Cyclic Alternating Pattern Sleep Database and achieved F_1-scores between 95% and 99% for 5 different sleep disorder categories. By contrast, Nasifoglu and Erogul [295] tested a fused CNN-ResNet approach for obstructive sleep apnea detection (accuracy 85.20%) and prediction (accuracy 82.30%) using data from a private database. Mukherjee et al [296] used a multilayer perceptron to detect sleep apnea from ECG recordings that originated from the Apnea-ECG Database, achieving an accuracy of 85.58%.

Other Clinical Analyses

Approximately 10.4% (24/230) of studies that applied DL methods to ECG data to perform other clinical analyses were identified during the literature search. Almost half (10/24, 42%) of the studies proposed a CNN approach, including Isasi et al [300], who used data from a private database to detect shockable and nonshockable rhythms during cardiopulmonary resuscitation with an accuracy of 96.10%, and Ozdemir et al [309], who used a private database to diagnose COVID-19 through ECG classification (accuracy 93.00%).

Other notable works include a study by Chang et al [311], which tested (sensitivity 84.60% and specificity 96.60%) an ECG12Net to detect digoxin toxicity by using private ECG signals from patients with digoxin toxicity and patients in the emergency room, and another study by Baghersalimi et al [313], which evaluated the performance (sensitivity 90.24% and specificity 91.58%) of a fused CNN-ResNet network to detect epileptic seizure events from single-lead ECG signals originating from a private database. Finally, Mazumder et al [318] implemented a CNN-LSTM structure for the detection of shockable rhythms in ECG signals from 2 public databases, achieving sensitivity scores between 94.68% and 99.21% and specificity scores between 92.77% and 99.68% for 2- and 8-second time windows, respectively.

Discussion

Principal Findings

DL has led to the creation of robust models that could potentially perform fast and reliable clinical diagnoses based on physiological signals. Remarkably, during the past 2 years, at least 230 studies that used DL on ECG data for various clinical applications were identified in the literature, which is a large number for such a short period, regardless of the application domain. This is mainly justified by the fact that DL methods can automatically capture distinctive features from ECG signals based on the trained models that achieve promising diagnostic performance, as shown in Multimedia Appendix 2 [93-322]. This constitutes a significant advantage compared with classical ML methods that perform manual feature selection and feature extraction—2 processes that conventionally require considerable effort and time [323]. Overall, CNN represents the most popular DL architecture and has been identified in most of the reviewed studies (142/230, 60.9% articles). On the contrary, 18.3% (42/230) of studies used LSTM architecture, whereas a ResNet architecture was used in 17.8% (41/230) of cases.

However, training a DL model is not always straightforward. Both architectural design choices and parameter tuning influence model performance; thus, multiple combinations must be considered. Furthermore, the training phase of DL algorithms typically involves complex computations that can be translated into long training times. This requires expensive state-of-the-art computer hardware, including graphics processing units that can dramatically accelerate the total execution time [324].

Another common problem with DL algorithms is overfitting; this occurs when the algorithm fits the noise and therefore performs well on the training set but fails to generalize its predictions to unseen data (ie, the testing set). For this reason, it is necessary to adopt an early stopping strategy during the training phase to prevent further training when the model's performance on unknown data starts to deteriorate. This is usually done by implementing a separate data set, called the validation set, which most of the time is a small percentage of the training set that is held back from training to provide an unbiased evaluation of the model during training. Moreover, random data splitting can introduce bias; thus, k-fold cross-validation or leave-one-out cross-validation strategies are preferred when training DL models. In addition, it is important that different sets (ie, training, validation, and testing) contain different patients, also known as interpatient data splitting, so that the study's results are more reliable. As concluded by this review and presented in Multimedia Appendix 2 [93-322], many researchers do not take this into consideration; hence, their results are questionable.

Another critical issue related to overfitting is the distribution of labels or predicted variables in the data set used for model development and validation. For instance, in the BP prediction problem, large stretches of constant BP from the same individual would bias the network toward a constant predictor with minimal error, with the network preferring to memorize patient-identifying features to predict the average BP for a patient rather than those which represent physiological metrics useful in predicting variable BP for the same patient. The resulting errors would be deceptively low if a patient's nominal BP does not change but, critically, would not be clinically useful in the setting of hypertensive or hypotensive crisis or to guide patient care. None of the assessed papers described the results, indicating that the predicted BP follows meaningful trends.

Recent attention in the medical field to the concept of BP variability [325] rather than clinical spot checks highlights the need for ambulatory BP monitors that are both ergonomic for the patient to increase compliance and comfort, as well as reliable and well validated. A common pitfall in the use of calibrated techniques is that subsequent test data points do not differ significantly from the calibration value and thus yield small errors in prediction, whereas the data are presented as an aggregate pooled correlation plot or Bland-Altman plot with a correlation value that simply reflects the range of BPs across the population rather than patient-specific BP variation [326,327]. In our review of articles using DL for BP prediction, we did not encounter significant attempts to address the issue

of BP variability in training data; in fact, many publications explicitly removed data points with hypertensive values or large pulse pressures from their data sets as “artifacts” [93-96,98].

In a calibration-less approach, a narrow range of variation would lead to a low prediction error even when predicting the population mean for each patient. If an ambulatory BP monitoring device plans to use AI-based techniques to measure variability, this variability must be represented in the training set for a model to learn to predict such changes adequately. A way of accomplishing this is to incorporate a variety of BP-modulating activities in the training data, which represent different sources of BP change and corresponding modulations in the feature space. For example, ice pressor tests may increase BP via peripheral vasoconstriction [328], whereas the valsalva maneuver increases chest pressure extrinsically [329] and may modulate input features such as heart rate in opposite ways, reducing the chance that bias-prone DL architectures learn misleading relationships.

In addition to the training and evaluation data, evaluation metrics and cost functions are areas with significant room for improvement. Mean squared error alone can be minimized with a constant predictor if the BP range does not vary significantly. Alternative cost functions such as cosine similarity, which is maximized with constant inputs, contrastive losses, or combinations thereof, have been successful in classification problems in imbalanced, rare event prediction problems such as critical events in patients with COVID-19 [330]. For other promising solutions, it would be prudent to examine similar trend prediction problems in other fields such as stock price movement, where progress has been made using intuitive data preparation and creative representation of the prediction targets, in this case, price changes, to generate trend deterministic predictions [331].

Furthermore, a vast majority of available ECG data sources experience data imbalance. This creates a major problem when trying to predict smaller classes that usually represent rare conditions or diseases that are as important as larger classes when designing health care decision support systems. To solve this problem, several oversampling techniques have been proposed, including random oversampling and undersampling, the synthetic minority oversampling technique [332], the adaptive synthetic sampling technique [333], the generative oversampling method [334], distribution-based balancing [335], and new loss functions such as focal loss [336], which can achieve both prediction error reduction and data imbalance handling. Papers addressing classification frequently use techniques to address class imbalance; however, evidence for such corrections in regression models does not appear as frequently or rigorously.

In addition, DL models are often characterized by black box behavior (lack of interpretability); that is, it is difficult for a human to understand why a particular result is generated by such complex architectures. This is crucial when training models for medical applications, as diagnoses based on unexplained model predictions are not usually accepted by medical experts. A possible solution to this problem is to take advantage of algorithms that are more easily interpretable, such as decision trees [337], additive models [338], attention-based networks [339], and sparse linear models [340], when designing a DL architecture. By contrast, several DL model interpretation approaches have been proposed in this direction, including permutation feature importance [341], partial dependence plots [342], and local interpretable model-agnostic explanations [343]. However, these techniques are rarely used in practice as they require additional time and effort. A useful technique that is used more often when dealing with medical images (and CNNs) is gradient-weighted class activation mapping [344], which makes CNN-based models more transparent by presenting visual explanations for their decisions.

Uncertainty quantification is another common problem associated with DL methods, which has recently drawn the attention of researchers. There are 2 main types of uncertainty: aleatoric (data uncertainty) and epistemic (knowledge uncertainty). It is important to evaluate the reliability and validity of DL methods before they can be tested in real-world applications; thus, uncertainty estimation should be provided. In the past few years, several uncertainty quantification techniques have been proposed, including deep Bayesian active learning [345], Monte Carlo dropout [346], Markov chain Monte Carlo [347], and Bayes by backprop [348].

Moreover, as presented in Multimedia Appendix 1 [54-92], there is no gold standard for data collection. As shown in Multimedia Appendix 2 [93-322], different studies used ECG data with distinct characteristics, namely, the number of leads, signal duration, and sample rate. In addition, many studies used multimodal data, such as photoplethysmograms, arterial BP, polysomnography, and electroencephalograms. Some studies used raw waveforms as input to their models, whereas others precomputed a set of features. This heterogeneity makes it difficult to compare study results; thus, finding the best algorithm is challenging, if not impossible.

Recent advancements [349] in materials and techniques to produce flexible, skin-integrated technology [350] have enabled the development of unique sensors and devices that can simultaneously measure both conventional and novel types of signals from the human body. Small wireless devices [351] such as these can extract continuous ECG; acceleration-based body orientation; physical activity [352]; vibrations such as heart sounds, breath sounds [353]; vocal processes [354]; and photoplethysmogram signals at multiple wavelengths and body locations. This wealth of physiological information that can be measured noninvasively and continuously throughout day-to-day life is potentially a treasure trove of useful insights into health status outside the rigidity of a clinical system. Tools such as DL have emerged as a tantalizing approach to take advantage of such multivariate data in the context of the increased complexity and unpredictability of ambulatory environments. With careful data curation and training approaches, as well as the use of intuitive, well-justified algorithms and network structures, explainable AI can also provide justifications for the use of novel features of underlying physiological relevance. Currently, the use of highly complex and computationally expensive DL models in wearable applications is limited. Generally, raw data are processed in a post hoc fashion after data have been uploaded to cloud servers, limiting real-time

feedback. However, recently, there have been developments by chip manufacturers to enable "edge inferencing" by bringing AI-enabling computational acceleration to the low–power-integrated circuit level, opening up the possibilities for low-latency applications of DL algorithms. We strongly believe that the creation of robust DL models that can assist medical experts in clinical decision-making is an important direction for future investigations.

In general, we believe that with this study, we (1) provided a complete and systematic account of the current state-of-the-art DL methods applied to ECG data; (2) identified several ECG data sources used in clinical diagnosis, even some not so widely cited databases; and (3) identified important open research problems and provided suggestions for future research directions in the field of DL and ECG data. Several important relevant review studies have already presented novel DL methods that are used on ECG data [355-357]. Nonetheless, none of them combine all the aforementioned characteristics, which makes this study innovative.

By contrast, the limitations of this study could be summarized as the fact that owing to the enormous number of studies focusing on DL and ECG data, we performed a review based only on articles that have been published in various journals between January 2020 and December 2021. Although the rationale behind this study was to identify all state-of-the-art DL methods that are applied to ECG data for various clinical applications, in the future, we intend to concentrate our efforts on providing a more complete account of DL methods that make good use of ECG data to address a specific clinical task (ie, congestive heart failure diagnosis).

Conclusions

In this study, we systematically reviewed 230 recently published articles on DL methods applied to ECG data for various clinical applications. We attempted to group the proposed DL approaches per field of application and summarize the most notable approaches among them. To the best of our knowledge, this is the first study that provides a complete account of the detailed strategy for designing each one of the proposed DL systems by recording the ECG data sources, data preprocessing techniques, model training, evaluation processes, and data splitting strategies that are implemented in each approach. Finally, open research problems and potential gaps were discussed to assess the future of the field and provide guidance to new researchers to design and implement reliable DL algorithms that can provide accurate diagnoses based on ECG data to support medical experts' efforts for clinical decision-making.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of the major electrocardiogram databases.
[DOCX File , 23 KB - medinform_v10i8e38454_app1.docx]

Multimedia Appendix 2
Summary of works carried out using deep-learning algorithms and electrocardiogram signals.
[DOCX File , 83 KB - medinform_v10i8e38454_app2.docx]

References

1. Schlant RC, Adolph RJ, DiMarco JP, Dreifus LS, Dunn MI, Fisch C, et al. Guidelines for electrocardiography. A report of the American college of cardiology/American heart association task force on assessment of diagnostic and therapeutic cardiovascular procedures (committee on electrocardiography). Circulation 1992 Mar;85(3):1221-1228 [FREE Full text] [doi: 10.1161/01.cir.85.3.1221] [Medline: 1537123]
2. Salerno SM, Alguire PC, Waxman HS, American College of Physicians. Training and competency evaluation for interpretation of 12-lead electrocardiograms: recommendations from the American College of Physicians. Ann Intern Med 2003 May 06;138(9):747-750 [FREE Full text] [doi: 10.7326/0003-4819-138-9-200305060-00012] [Medline: 12729430]
3. Cook DA, Oh SY, Pusic MV. Accuracy of physicians' electrocardiogram interpretations: a systematic review and meta-analysis. JAMA Intern Med 2020 Nov 01;180(11):1461-1471 [FREE Full text] [doi: 10.1001/jamainternmed.2020.3989] [Medline: 32986084]
4. Ramesh AN, Kambhampati C, Monson JR, Drew PJ. Artificial intelligence in medicine. Ann R Coll Surg Engl 2004 Sep;86(5):334-338 [FREE Full text] [doi: 10.1308/147870804290] [Medline: 15333167]
5. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018 Jun 12;71(23):2668-2679 [FREE Full text] [doi: 10.1016/j.jacc.2018.03.521] [Medline: 29880128]
6. Awan SE, Sohel F, Sanfilippo FM, Bennamoun M, Dwivedi G. Machine learning in heart failure: ready for prime time. Curr Opin Cardiol 2018 Mar;33(2):190-195 [FREE Full text] [doi: 10.1097/HCO.0000000000000491] [Medline: 29194052]
7. Lai Y. A comparison of traditional machine learning and deep learning in image recognition. J Phys Conf Ser 2019 Oct 01;1314(1):012148 [FREE Full text] [doi: 10.1088/1742-6596/1314/1/012148]

8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444 [FREE Full text] [doi: 10.1038/nature14539] [Medline: 26017442]
9. Deep learning techniques: an overview. In: Advanced Machine Learning Technologies and Applications. Singapore: Springer; 2021. URL: https://link.springer.com/chapter/10.1007/978-981-15-3383-9_54
10. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989 Dec;1(4):541-551. [doi: 10.1162/neco.1989.1.4.541]
11. Object recognition with gradient-based learning. In: Shape, Contour and Grouping in Computer Vision. Berlin, Heidelberg: Springer; 1999.
12. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018 Oct 08;42(11):226 [FREE Full text] [doi: 10.1007/s10916-018-1088-1] [Medline: 30298337]
13. Kamaleswaran R, Mahajan R, Akbilgic O. A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using single lead electrocardiograms of variable length. Physiol Meas 2018 Mar 27;39(3):035006 [FREE Full text] [doi: 10.1088/1361-6579/aaaa9d] [Medline: 29369044]
14. Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 2019 Jun;16(3):031001 [FREE Full text] [doi: 10.1088/1741-2552/ab0ab5] [Medline: 30808014]
15. Bardou D, Zhang K, Ahmad SM. Lung sounds classification using convolutional neural networks. Artif Intell Med 2018 Jun;88:58-69 [FREE Full text] [doi: 10.1016/j.artmed.2018.04.008] [Medline: 29724435]
16. Zhang Q, Fu L, Gu L. A cascaded convolutional neural network for assessing signal quality of dynamic ECG. Comput Math Methods Med 2019;2019:7095137 [FREE Full text] [doi: 10.1155/2019/7095137] [Medline: 31781289]
17. Wu H, Soraghan J, Lowit A, Di Caterina G. Convolutional neural networks for pathological voice detection. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:1-4 [FREE Full text] [doi: 10.1109/EMBC.2018.8513222] [Medline: 30440307]
18. Chriskos P, Frantzidis CA, Gkivogkli PT, Bamidis PD, Kourtidou-Papadeli C. Automatic sleep staging employing convolutional neural networks and cortical connectivity images. IEEE Trans Neural Netw Learning Syst 2020 Jan;31(1):113-123 [FREE Full text] [doi: 10.1109/tnnls.2019.2899781]
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV, USA URL: https://doi.org/10.1109/cvpr.2016.90 [doi: 10.1109/cvpr.2016.90]
20. Xu H, Baxter JS, Akin O, Cantor-Rivera D. Prostate cancer detection using residual networks. Int J Comput Assist Radiol Surg 2019 Oct;14(10):1647-1650 [FREE Full text] [doi: 10.1007/s11548-019-01967-5] [Medline: 30972686]
21. Wang EK, Zhang X, Pan L, Cheng C, Dimitrakopoulou-Strauss A, Li Y, et al. Multi-path dilated residual network for nuclei segmentation and detection. Cells 2019 May 23;8(5):499 [FREE Full text] [doi: 10.3390/cells8050499] [Medline: 31126166]
22. Datong C, Minghui L, Cheng J, Yue S, Dongbin X, Yueming L. Coronary calcium detection based on improved deep residual network in mimics. J Med Syst 2019 Mar 25;43(5):119 [FREE Full text] [doi: 10.1007/s10916-019-1218-4] [Medline: 30911850]
23. Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. Int J Comput Assist Radiol Surg 2017 Oct;12(10):1799-1808 [FREE Full text] [doi: 10.1007/s11548-017-1605-6] [Medline: 28501942]
24. Usama M, Ahmad B, Xiao W, Hossain MS, Muhammad G. Self-attention based recurrent convolutional neural network for disease prediction using healthcare data. Comput Methods Programs Biomed 2020 Jul;190:105191 [FREE Full text] [doi: 10.1016/j.cmpb.2019.105191] [Medline: 31753591]
25. Chakravarty A, Sivaswamy J. RACE-Net: a recurrent neural network for biomedical image segmentation. IEEE J Biomed Health Inform 2019 May;23(3):1151-1162 [FREE Full text] [doi: 10.1109/jbhi.2018.2852635]
26. Arsenali B, van Dijk J, Ouweltjes O, den Brinker B, Pevernagie D, Krijn R, et al. Recurrent neural network for classification of snoring and non-snoring sound events. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:328-331 [FREE Full text] [doi: 10.1109/EMBC.2018.8512251] [Medline: 30440404]
27. Rajeev R, Samath JA, Karthikeyan NK. An intelligent recurrent neural network with long short-term memory (LSTM) BASED batch normalization for medical image denoising. J Med Syst 2019 Jun 15;43(8):234 [FREE Full text] [doi: 10.1007/s10916-019-1371-9] [Medline: 31203556]
28. Liu M, Cheng D, Yan W, Alzheimer's Disease Neuroimaging Initiative. Classification of alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. Front Neuroinform 2018;12:35 [FREE Full text] [doi: 10.3389/fninf.2018.00035] [Medline: 29970996]
29. Beeksma M, Verberne S, van den Bosch A, Das E, Hendrickx I, Groenewoud S. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. BMC Med Inform Decis Mak 2019 Feb 28;19(1):36 [FREE Full text] [doi: 10.1186/s12911-019-0775-2] [Medline: 30819172]
30. Gao J, Zhang H, Lu P, Wang Z. An effective LSTM recurrent network to detect arrhythmia on imbalanced ECG dataset. J Healthc Eng 2019;2019:6320651 [FREE Full text] [doi: 10.1155/2019/6320651] [Medline: 31737240]
31. Ahmedt-Aristizabal D, Fookes C, Nguyen K, Sridharan S. Deep classification of epileptic signals. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:332-335 [FREE Full text] [doi: 10.1109/EMBC.2018.8512249] [Medline: 30440405]

32. Wollmann T, Gunkel M, Chung I, Erfle H, Rippe K, Rohr K. GRUU-Net: integrated convolutional and gated recurrent neural network for cell segmentation. Med Image Anal 2019 Aug;56:68-79 [FREE Full text] [doi: 10.1016/j.media.2019.04.011] [Medline: 31200289]
33. Taheri Dezaki F, Liao Z, Luong C, Girgis H, Dhungel N, Abdi AH, et al. Cardiac phase detection in echocardiograms with densely gated recurrent neural networks and global extrema loss. IEEE Trans Med Imaging 2019 Aug;38(8):1821-1832 [FREE Full text] [doi: 10.1109/tmi.2018.2888807]
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. ArXiv 2017.
35. Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, et al. A wide and deep transformer neural network for 12-lead ECG classification. Computing Cardiol 2020;47 [FREE Full text] [doi: 10.22489/cinc.2020.107]
36. Wang L, Niu D, Zhao X, Wang X, Hao M, Che H. A comparative analysis of novel deep learning and ensemble learning models to predict the allergenicity of food proteins. Foods 2021 Apr 09;10(4):809 [FREE Full text] [doi: 10.3390/foods10040809] [Medline: 33918556]
37. Yang F, Wang X, Ma H, Li J. Transformers-sklearn: a toolkit for medical language understanding with transformer-based models. BMC Med Inform Decis Mak 2021 Jul 30;21(Suppl 2):90 [FREE Full text] [doi: 10.1186/s12911-021-01459-0] [Medline: 34330244]
38. Rajan K, Zielesny A, Steinbeck C. DECIMER 1.0: deep learning for chemical image recognition using transformers. J Cheminform 2021 Aug 17;13(1):61 [FREE Full text] [doi: 10.1186/s13321-021-00538-8] [Medline: 34404468]
39. Zhu Y, Wang MD, Tong L, Deshpande SR. Improved prediction on heart transplant rejection using convolutional autoencoder and multiple instance learning on whole-slide imaging. IEEE EMBS Int Conf Biomed Health Inform 2019 May;2019 [FREE Full text] [doi: 10.1109/bhi.2019.8834632] [Medline: 32577622]
40. Song T, Sanchez V, EI Daly H, Rajpoot NM. Simultaneous cell detection and classification in bone marrow histology images. IEEE J Biomed Health Inform 2019 Jul;23(4):1469-1476 [FREE Full text] [doi: 10.1109/jbhi.2018.2878945]
41. Xu X, Gu H, Wang Y, Wang J, Qin P. Autoencoder based feature selection method for classification of anticancer drug response. Front Genet 2019;10:233 [FREE Full text] [doi: 10.3389/fgene.2019.00233] [Medline: 30972101]
42. Gordon M, Williams C. PVC detection using a convolutional autoencoder and random forest classifier. Biocomputing 2019:42-53 [FREE Full text] [doi: 10.1142/9789813279827_0005]
43. Tong L, Wu H, Wang MD. CAESNet: convolutional AutoEncoder based Semi-supervised Network for improving multiclass classification of endomicroscopic images. J Am Med Inform Assoc 2019 Nov 01;26(11):1286-1296 [FREE Full text] [doi: 10.1093/jamia/ocz089] [Medline: 31260038]
44. Vaillant R. Original approach for the localisation of objects in images. IEE Proc Vis Image Process 1994;141(4):245 [FREE Full text] [doi: 10.1049/ip-vis:19941301]
45. Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: a survey. Mech Syst Signal Process 2021 Apr;151:107398 [FREE Full text] [doi: 10.1016/j.ymssp.2020.107398]
46. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct;323(6088):533-536 [FREE Full text] [doi: 10.1038/323533a0]
47. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780 [FREE Full text] [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
48. Cho K, Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. 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); Oct, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1179]
49. Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations. Cambridge, Massachusetts, United States: MIT Press; 1986.
50. Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. 2008 Presented at: ICML '08: The 25th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming; Jul 5 - 9, 2008; Helsinki Finland URL: https://doi.org/10.1145/1390156.1390294 [doi: 10.1145/1390156.1390294]
51. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv 2014 [FREE Full text]
52. Li F, Qiao H, Zhang B. Discriminatively boosted image clustering with fully convolutional auto-encoders. Pattern Recognition 2018 Nov;83:161-173 [FREE Full text] [doi: 10.1016/j.patcog.2018.05.019]
53. Welcome to VOSviewer. VOSviewer. URL: https://www.vosviewer.com/ [accessed 2021-11-09]
54. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220 [FREE Full text] [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
55. Silva I, Behar J, Sameni R, Zhu T, Oster J, Clifford GD, et al. Noninvasive fetal ECG: the PhysioNet/Computing in cardiology challenge 2013. Comput Cardiol (2010) 2013 Mar;40:149-152 [FREE Full text] [Medline: 25401167]
56. Clifford GD, Silva I, Moody B, Li Q, Kella D, Shahin A, et al. The PhysioNet/Computing in Cardiology Challenge 2015: reducing false arrhythmia alarms in the ICU. In: Proceedings of the 2015 Computing in Cardiology Conference (CinC).

2015 Presented at: 2015 Computing in Cardiology Conference (CinC); Sep 06-09, 2015; Nice, France URL: https://doi.org/10.1109/cic.2015.7408639 [doi: 10.1109/cic.2015.7408639]
57. Clifford GD, Liu C, Moody B, Lehman L, Silva I, Li Q, et al. AF classification from a short single lead ECG recording: the physionet/computing in cardiology challenge 2017. Comput Cardiol 2017;44 [FREE Full text] [doi: 10.22489/cinc.2017.065-469]
58. Ghassemi MM, Moody BE, Lehman LW, Song C, Li Q, Sun H, et al. You snooze, you win: the PhysioNet/computing in cardiology challenge 2018. Comput Cardiol (2010) 2018 Sep;45 [FREE Full text] [doi: 10.22489/cinc.2018.049] [Medline: 34796237]
59. Perez Alday EA, Gu A, J Shah A, Robichaux C, Ian Wong A, Liu C, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas 2021 Jan 01;41(12):124003 [FREE Full text] [doi: 10.1088/1361-6579/abc960] [Medline: 33176294]
60. Jezewski J, Matonia A, Kupka T, Roj D, Czabanski R. Determination of fetal heart rate from abdominal signals: evaluation of beat-to-beat accuracy in relation to the direct fetal electrocardiogram. Biomed Tech (Berl) 2012 Oct;57(5):383-394 [FREE Full text] [doi: 10.1515/bmt-2011-0130] [Medline: 25854665]
61. Moody GE. Spontaneous termination of atrial fibrillation: a challenge from physionet and computers in cardiology 2004. In: Proceedings of the Computers in Cardiology, 2004. 2004 Presented at: Computers in Cardiology, 2004; Sep 19-22, 2004; Chicago, IL, USA. [doi: 10.1109/cic.2004.1442881]
62. Penzel T, Moody G, Mark RG, Goldberger AL, Peter JH. The apnea-ECG database. In: Proceedings of the Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163). 2000 Presented at: Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163); Sep 24-27, 2000; Cambridge, MA, USA URL: https://doi.org/10.1109/cic.2000.898505 [doi: 10.1109/cic.2000.898505]
63. Baim DS, Colucci WS, Monrad ES, Smith HS, Wright RF, Lanoue A, et al. Survival of patients with severe congestive heart failure treated with oral milrinone. J Am Coll Cardiol 1986 Mar;7(3):661-670 [FREE Full text] [doi: 10.1016/s0735-1097(86)80478-8]
64. Terzano MG, Parrino L, Sherieri A, Chervin R, Chokroverty S, Guilleminault C, et al. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med 2001 Nov;2(6):537-553 [FREE Full text] [doi: 10.1016/s1389-9457(01)00149-6]
65. Zheng J, Fu G, Anderson K, Chu H, Rakovski C. A 12-Lead ECG database to identify origins of idiopathic ventricular arrhythmia containing 334 patients. Sci Data 2020 Mar 23;7(1):98 [FREE Full text] [doi: 10.1038/s41597-020-0440-8] [Medline: 32251335]
66. Zheng J, Chu H, Struppa D, Zhang J, Yacoub SM, El-Askary H, et al. Optimal multi-stage arrhythmia classification approach. Sci Rep 2020 Feb 19;10(1):2898 [FREE Full text] [doi: 10.1038/s41598-020-59821-7] [Medline: 32076033]
67. Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Sci Data 2020 Feb 12;7(1):48 [FREE Full text] [doi: 10.1038/s41597-020-0386-x] [Medline: 32051412]
68. Da Silva HP, Lourenço A, Fred A, Raposo N, Aires-de-Sousa M. Check your biosignals here: a new dataset for off-the-person ECG biometrics. Comput Methods Programs Biomed 2014 Feb;113(2):503-514 [FREE Full text] [doi: 10.1016/j.cmpb.2013.11.017] [Medline: 24377903]
69. Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. J Med Imaging Health Inform 2018 Sep 01;8(7):1368-1373 [FREE Full text] [doi: 10.1166/jmihi.2018.2442]
70. Gao H, Liu C, Wang X, Zhao L, Shen Q, Ng EY, et al. An open-access ECG database for algorithm evaluation of QRS detection and heart rate estimation. J Med Imaging Health Inform 2019 Dec 01;9(9):1853-1858 [FREE Full text] [doi: 10.1166/jmihi.2019.2800]
71. Cai Z, Liu C, Gao H, Wang X, Zhao L, Shen Q, et al. An open-access long-term wearable ECG database for premature ventricular contractions and supraventricular premature beat detection. J Med Imaging Health Inform 2020 Nov 01;10(11):2663-2667 [FREE Full text] [doi: 10.1166/jmihi.2020.32892663]
72. Nolle FM. CREI-GARD: a new concept in computerized arrhythmia monitoring systems. Comput Cardiol 1986;13:515-518.
73. Biometric human identification based on ECG. Saint-Petersburg, Russian Federation. URL: https://archive.physionet.org/physiobank/database/ecgiddb/images/ [accessed 2022-07-26]
74. Iyengar N, Peng CK, Morin R, Goldberger AL, Lipsitz LA. Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. Am J Physiol Regul Integr Comp Physiol 1996 Oct 01;271(4):R1078-R1084. [doi: 10.1152/ajpregu.1996.271.4.r1078]
75. Petrutiu S, Sahakian A, Swiryn S. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. Europace 2007 Jul;9(7):466-470. [doi: 10.1093/europace/eum096] [Medline: 17540663]
76. Moody GB, Mark RG. A database to support development and evaluation of intelligent intensive care monitoring. In: Proceedings of the Computers in Cardiology 1996. 1996 Presented at: Computers in Cardiology 1996; Sep 8-11, 1996; Indianapolis, IN, USA URL: https://doi.org/10.1109/cic.1996.542622 [doi: 10.1109/cic.1996.542622]

77. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database*. Crit Care Med 2011;39(5):952-960 [FREE Full text] [doi: 10.1097/ccm.0b013e31820a92c6]
78. MIMIC-III waveform database (version 1.0). PhysioNet. URL: https://doi.org/10.13026/c2607m [accessed 2022-07-26]
79. 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(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
80. Moody G, Mark R. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001;20(3):45-50. [doi: 10.1109/51.932724] [Medline: 11446209]
81. Moody GB, Mark RG. A new method for detecting atrial fibrillation using R-R intervals. Comput Cardiol 1983:227.
82. The development and analysis of a ventricular fibrillation detector. Massachusetts Institute of Technology. 1986. URL: https://dspace.mit.edu/handle/1721.1/92988 [accessed 2022-07-26]
83. Moody GB, Muldrow WE, Mark RG. A noise stress test for arrhythmia detectors. Comput Cardiol 1984:381-384.
84. ST Segment Characterization for Long Term Automated ECG Analysis. Cambridge: Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science; 1983.
85. Moody G, Goldberger AL, McClennen S, Swiryn SP. Predicting the onset of paroxysmal atrial fibrillation: the Computers in Cardiology Challenge 2001. In: Proceedings of the Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287). 2001 Presented at: Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287); Sep 23-26, 2001; Rotterdam, Netherlands. [doi: 10.1109/cic.2001.977604]
86. Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das internet. De Gruyter 1995;40(s1):317-318 [FREE Full text] [doi: 10.1515/bmte.1995.40.s1.317]
87. Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020 May 25;7(1):154 [FREE Full text] [doi: 10.1038/s41597-020-0495-6] [Medline: 32451379]
88. Laguna P, Mark RG, Goldberg A, Moody GB. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In: Proceedings of the Computers in Cardiology 1997. 1997 Presented at: Computers in Cardiology 1997; Sep 07-10, 1997; Lund, Sweden URL: https://doi.org/10.1109/cic.1997.648140 [doi: 10.1109/cic.1997.648140]
89. Melillo P, Izzo R, Orrico A, Scala P, Attanasio M, Mirra M, et al. Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. PLoS One 2015;10(3):e0118504 [FREE Full text] [doi: 10.1371/journal.pone.0118504] [Medline: 25793605]
90. St. Vincent's University Hospital / University College Dublin Sleep Apnea Database. PhysioNet. URL: https://physionet.org/content/ucddb/1.0.0/ [accessed 2022-07-26]
91. Khamis H, Weiss R, Xie Y, Chang C, Lovell NH, Redmond SJ. QRS detection algorithm for telehealth electrocardiogram recordings. IEEE Trans Biomed Eng 2016 Jul;63(7):1377-1388 [FREE Full text] [doi: 10.1109/tbme.2016.2549060]
92. Bizzego A, Gabrieli G, Furlanello C, Esposito G. Comparison of wearable and clinical devices for acquisition of peripheral nervous system signals. Sensors (Basel) 2020 Nov 27;20(23):6778 [FREE Full text] [doi: 10.3390/s20236778] [Medline: 33260880]
93. Li YH, Harfiya LN, Purwandari K, Lin YD. Real-time cuffless continuous blood pressure estimation using deep learning model. Sensors (Basel) 2020 Sep 30;20(19):5606 [FREE Full text] [doi: 10.3390/s20195606] [Medline: 33007891]
94. Miao F, Wen B, Hu Z, Fortino G, Wang XP, Liu ZD, et al. Continuous blood pressure measurement from one-channel electrocardiogram signal using deep-learning techniques. Artif Intell Med 2020 Aug;108:101919 [FREE Full text] [doi: 10.1016/j.artmed.2020.101919] [Medline: 32972654]
95. Hill BL, Rakocz N, Rudas Á, Chiang JN, Wang S, Hofer I, et al. Imputation of the continuous arterial line blood pressure waveform from non-invasive measurements using deep learning. Sci Rep 2021 Aug 03;11(1):15755 [FREE Full text] [doi: 10.1038/s41598-021-94913-y] [Medline: 34344934]
96. Paviglianiti A, Randazzo V, Villata S, Cirrincione G, Pasero E. A comparison of deep learning techniques for arterial blood pressure prediction. Cognit Comput 2021 Aug 27:1-22 [FREE Full text] [doi: 10.1007/s12559-021-09910-0] [Medline: 34466163]
97. Fan X, Wang H, Zhao Y, Li Y, Tsui KL. An adaptive weight learning-based multitask deep network for continuous blood pressure estimation using electrocardiogram signals. Sensors (Basel) 2021 Feb 25;21(5):1595 [FREE Full text] [doi: 10.3390/s21051595] [Medline: 33668778]
98. Jeong DU, Lim KM. Combined deep CNN-LSTM network-based multitasking learning architecture for noninvasive continuous blood pressure estimation using difference in ECG-PPG features. Sci Rep 2021 Jun 29;11(1):13539 [FREE Full text] [doi: 10.1038/s41598-021-92997-0] [Medline: 34188132]
99. Baalman SW, Schroevers FE, Oakley AJ, Brouwer TF, van der Stuijt W, Bleijendaal H, et al. A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples. Int J Cardiol 2020 Oct 01;316:130-136 [FREE Full text] [doi: 10.1016/j.ijcard.2020.04.046] [Medline: 32315684]

100. Gao Y, Wang H, Liu Z. An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss. Knowledge Based Systems 2021 Jan;212:106589 [FREE Full text] [doi: 10.1016/j.knosys.2020.106589]
101. Cai W, Chen Y, Guo J, Han B, Shi Y, Ji L, et al. Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. Comput Biol Med 2020 Jan;116:103378 [FREE Full text] [doi: 10.1016/j.compbiomed.2019.103378] [Medline: 31778896]
102. Shi H, Wang H, Qin C, Zhao L, Liu C. An incremental learning system for atrial fibrillation detection based on transfer learning and active learning. Comput Methods Programs Biomed 2020 Apr;187:105219 [FREE Full text] [doi: 10.1016/j.cmpb.2019.105219] [Medline: 31786450]
103. Ghosh SK, Tripathy RK, Paternina MR, Arrieta JJ, Zamora-Mendez A, Naik GR. Detection of atrial fibrillation from single lead ECG signal using multirate cosine filter bank and deep neural network. J Med Syst 2020 May 10;44(6):114 [FREE Full text] [doi: 10.1007/s10916-020-01565-y] [Medline: 32388733]
104. Hsieh CH, Li YS, Hwang BJ, Hsiao CH. Detection of atrial fibrillation using 1D convolutional neural network. Sensors (Basel) 2020 Apr 10;20(7):2136 [FREE Full text] [doi: 10.3390/s20072136] [Medline: 32290113]
105. Mousavi S, Afghah F, Acharya UR. HAN-ECG: an interpretable atrial fibrillation detection model using hierarchical attention networks. Comput Biol Med 2020 Dec;127:104057 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.104057] [Medline: 33126126]
106. Tran L, Li Y, Nocera L, Shahabi C, Xiong L. MultiFusionNet: atrial fibrillation detection with deep neural networks. AMIA Jt Summits Transl Sci Proc 2020;2020:654-663 [FREE Full text] [Medline: 32477688]
107. Abdelazez M, Rajan S, Chan AD. Transfer learning for detection of atrial fibrillation in deterministic compressive sensed ECG. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:5398-5401 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9175813] [Medline: 33019201]
108. Buscema PM, Grossi E, Massini G, Breda M, Della Torre F. Computer Aided Diagnosis for atrial fibrillation based on new artificial adaptive systems. Comput Methods Programs Biomed 2020 Jul;191:105401 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105401] [Medline: 32146212]
109. Oster J, Hopewell JC, Ziberna K, Wijesurendra R, Camm CF, Casadei B, et al. Identification of patients with atrial fibrillation: a big data exploratory analysis of the UK Biobank. Physiol Meas 2020 Mar 06;41(2):025001 [FREE Full text] [doi: 10.1088/1361-6579/ab6f9a] [Medline: 31978903]
110. Lai D, Bu Y, Su Y, Zhang X, Ma C. Non-standardized patch-based ECG lead together with deep learning based algorithm for automatic screening of atrial fibrillation. IEEE J Biomed Health Inform 2020 Jun;24(6):1569-1578 [FREE Full text] [doi: 10.1109/jbhi.2020.2980454]
111. Kwon J, Cho Y, Jeon K, Cho S, Kim K, Baek SD, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. Lancet Digital Health 2020 Jul;2(7):e358-e367 [FREE Full text] [doi: 10.1016/s2589-7500(20)30108-4]
112. Kwon J, Lee SY, Jeon K, Lee Y, Kim K, Park J, et al. Deep learning–based algorithm for detecting aortic stenosis using electrocardiography. J Am Heart Assoc 2020 Apr 09;9(7):e014717 [FREE Full text] [doi: 10.1161/jaha.119.014717]
113. Hsu PY, Cheng CK. Arrhythmia classification using deep learning and machine learning with features extracted from waveform-based signal processing. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:292-295 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9176679] [Medline: 33017986]
114. Chen TM, Huang CH, Shih ES, Hu YF, Hwang MJ. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. iScience 2020 Mar 27;23(3):100886 [FREE Full text] [doi: 10.1016/j.isci.2020.100886] [Medline: 32062420]
115. Cheng Y, Ye Y, Hou M, He W, Pan T. Multi-label arrhythmia classification from fixed-length compressed ECG segments in real-time wearable ECG monitoring. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:580-583 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9176188] [Medline: 33018055]
116. Lennox C, Mahmud MS. Robust classification of cardiac arrhythmia using a deep neural network. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:288-291 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9175213] [Medline: 33017985]
117. Chen M, Wang G, Ding Z, Li J, Yang H. Unsupervised domain adaptation for ECG arrhythmia classification. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:304-307 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9175928] [Medline: 33017989]
118. Wang D, Meng Q, Chen D, Zhang H, Xu L. Automatic detection of arrhythmia based on multi-resolution representation of ECG signal. Sensors (Basel) 2020 Mar 12;20(6):1579 [FREE Full text] [doi: 10.3390/s20061579] [Medline: 32178296]
119. Liang Y, Yin S, Tang Q, Zheng Z, Elgendi M, Chen Z. Deep learning algorithm classifies heartbeat events based on electrocardiogram signals. Front Physiol 2020;11:569050 [FREE Full text] [doi: 10.3389/fphys.2020.569050] [Medline: 33117191]
120. Wang R, Fan J, Li Y. Deep multi-scale fusion neural network for multi-class arrhythmia detection. IEEE J Biomed Health Inform 2020 Sep;24(9):2461-2472 [FREE Full text] [doi: 10.1109/jbhi.2020.2981526]

121. Zhang J, Liu A, Gao M, Chen X, Zhang X, Chen X. ECG-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network. Artif Intell Med 2020 Jun;106:101856 [FREE Full text] [doi: 10.1016/j.artmed.2020.101856] [Medline: 32593390]
122. Sanjana K, Sowmya V, Gopalakrishnan EA, Soman KP. Explainable artificial intelligence for heart rate variability in ECG signal. Healthc Technol Lett 2020 Dec;7(6):146-154 [FREE Full text] [doi: 10.1049/htl.2020.0033] [Medline: 33425369]
123. Hata E, Seo C, Nakayama M, Iwasaki K, Ohkawauchi T, Ohya J. Classification of aortic stenosis using ECG by deep learning and its analysis using grad-CAM. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:1548-1551 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9175151] [Medline: 33018287]
124. Hu J, Zhao W, Jia D, Yan C, Wang H, Li Z, et al. Deep multi-instance networks for bundle branch block detection from multi-lead ECG. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:353-356 [FREE Full text] [doi: 10.1109/EMBC44109.2020.9175909] [Medline: 33018001]
125. Butun E, Yildirim O, Talo M, Tan R, Rajendra Acharya U. 1D-CADCapsNet: one dimensional deep capsule networks for coronary artery disease detection using ECG signals. Phys Med 2020 Feb;70:39-48 [FREE Full text] [doi: 10.1016/j.ejmp.2020.01.007] [Medline: 31962284]
126. Kwon JM, Kim KH, Jeon KH, Lee SY, Park J, Oh BH. Artificial intelligence algorithm for predicting cardiac arrest using electrocardiography. Scand J Trauma Resusc Emerg Med 2020 Oct 06;28(1):98 [FREE Full text] [doi: 10.1186/s13049-020-00791-0] [Medline: 33023615]
127. Yildirim O, Talo M, Ciaccio EJ, Tan RS, Acharya UR. Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ECG records. Comput Methods Programs Biomed 2020 Dec;197:105740 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105740] [Medline: 32932129]
128. Zhang X, Gu K, Miao S, Zhang X, Yin Y, Wan C, et al. Automated detection of cardiovascular disease by electrocardiogram signal analysis: a deep learning system. Cardiovasc Diagn Ther 2020 Apr;10(2):227-235 [FREE Full text] [doi: 10.21037/cdt.2019.12.10] [Medline: 32420103]
129. Thiagarajan JJ, Rajan D, Katoch S, Spanias A. DDxNet: a deep learning model for automatic interpretation of electronic health records, electrocardiograms and electroencephalograms. Sci Rep 2020 Oct 02;10(1):16428 [FREE Full text] [doi: 10.1038/s41598-020-73126-9] [Medline: 33009423]
130. Lin C, Lin C, Fang W, Hsu C, Chen S, Huang K, et al. A deep-learning algorithm (ECG12Net) for detecting hypokalemia and hyperkalemia by electrocardiography: algorithm development. JMIR Med Inform 2020 Mar 05;8(3):e15931 [FREE Full text] [doi: 10.2196/15931] [Medline: 32134388]
131. Jeon E, Oh K, Kwon S, Son H, Yun Y, Jung E, et al. A lightweight deep learning model for fast electrocardiographic beats classification with a wearable cardiac monitor: development and validation study. JMIR Med Inform 2020 Mar 12;8(3):e17037 [FREE Full text] [doi: 10.2196/17037] [Medline: 32163037]
132. Niu L, Chen C, Liu H, Zhou S, Shu M. A deep-learning approach to ECG classification based on adversarial domain adaptation. Healthcare (Basel) 2020 Oct 27;8(4):437 [FREE Full text] [doi: 10.3390/healthcare8040437] [Medline: 33121038]
133. Rincon JA, Guerra-Ojeda S, Carrascosa C, Julian V. An iot and fog computing-based monitoring system for cardiovascular patients with automatic ECG classification using deep neural networks. Sensors (Basel) 2020 Dec 21;20(24):7353 [FREE Full text] [doi: 10.3390/s20247353] [Medline: 33371514]
134. van de Leur R, Blom L, Gavves E, Hof I, van der Heijden J, Clappers N, et al. Automatic triage of 12‐lead ECGs using deep convolutional neural networks. J Am Heart Assoc 2020 May 18;9(10):e015138 [FREE Full text] [doi: 10.1161/jaha.119.015138]
135. Niu J, Tang Y, Sun Z, Zhang W. Inter-patient ECG classification with symbolic representations and multi-perspective convolutional neural networks. IEEE J Biomed Health Inform 2020 May;24(5):1321-1332 [FREE Full text] [doi: 10.1109/jbhi.2019.2942938]
136. Saadatnejad S, Oveisi M, Hashemi M. LSTM-based ECG classification for continuous monitoring on personal wearable devices. IEEE J Biomed Health Inform 2020 Feb;24(2):515-523 [FREE Full text] [doi: 10.1109/jbhi.2019.2911367]
137. Van Steenkiste G, van Loon G, Crevecoeur G. Transfer learning in ECG classification from human to horse using a novel parallel neural network architecture. Sci Rep 2020 Jan 13;10(1):186 [FREE Full text] [doi: 10.1038/s41598-019-57025-2] [Medline: 31932667]
138. Liu H, Zhao Z, Chen X, Yu R, She Q. Using the VQ-VAE to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records. Comput Methods Programs Biomed 2020 Nov;196:105639 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105639] [Medline: 32674047]
139. Vijayarangan S, Murugesan B, Vignesh R, Preejith SP, Joseph J, Sivaprakasam M. Interpreting deep neural networks for single-lead ECG arrhythmia classification. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 20-24, 2020; Montreal, QC, Canada URL: https://doi.org/10.1109/embc44109.2020.9176396 [doi: 10.1109/embc44109.2020.9176396]
140. Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, et al. Author Correction: automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020 May 01;11(1):2227 [FREE Full text] [doi: 10.1038/s41467-020-16172-1] [Medline: 32358526]

141. Zhu H, Cheng C, Yin H, Li X, Zuo P, Ding J, et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. Lancet Digital Health 2020 Jul;2(7):e348-e357 [FREE Full text] [doi: 10.1016/s2589-7500(20)30107-2]
142. Lih OS, Jahmunah V, San TR, Ciaccio EJ, Yamakawa T, Tanabe M, et al. Comprehensive electrocardiographic diagnosis based on deep learning. Artif Intell Med 2020 Mar;103:101789 [FREE Full text] [doi: 10.1016/j.artmed.2019.101789] [Medline: 32143796]
143. Mousavi S, Fotoohinasab A, Afghah F. Single-modal and multi-modal false arrhythmia alarm reduction using attention-based convolutional and recurrent neural networks. PLoS One 2020;15(1):e0226990 [FREE Full text] [doi: 10.1371/journal.pone.0226990] [Medline: 31923226]
144. Shahin M, Oo E, Ahmed B. Adversarial multi-task learning for robust end-to-end ECG-based heartbeat classification. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Presented at: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 20-24, 2020; Montreal, QC, Canada URL: https://doi.org/10.1109/embc44109.2020.9175640 [doi: 10.1109/embc44109.2020.9175640]
145. Romdhane TF, Alhichri H, Ouni R, Atri M. Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss. Comput Biol Med 2020 Aug;123:103866 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103866] [Medline: 32658786]
146. Li Z, Zhou D, Wan L, Li J, Mou W. Heartbeat classification using deep residual convolutional neural network from 2-lead electrocardiogram. J Electrocardiol 2020;58:105-112 [FREE Full text] [doi: 10.1016/j.jelectrocard.2019.11.046] [Medline: 31812617]
147. Soh DC, Ng EY, Jahmunah V, Oh SL, Tan RS, Acharya UR. Automated diagnostic tool for hypertension using convolutional neural network. Comput Biol Med 2020 Nov;126:103999 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103999] [Medline: 32992139]
148. Porumb M, Stranges S, Pescapè A, Pecchia L. Precision medicine and artificial intelligence: a pilot study on deep learning for hypoglycemic events detection based on ECG. Sci Rep 2020 Jan 13;10(1):170 [FREE Full text] [doi: 10.1038/s41598-019-56927-5] [Medline: 31932608]
149. Kwon JM, Jeon KH, Kim HM, Kim MJ, Lim SM, Kim KH, et al. Comparing the performance of artificial intelligence and conventional diagnosis criteria for detecting left ventricular hypertrophy using electrocardiography. Europace 2020 Mar 01;22(3):412-419 [FREE Full text] [doi: 10.1093/europace/euz324] [Medline: 31800031]
150. Cho Y, Kwon JM, Kim KH, Medina-Inojosa JR, Jeon KH, Cho S, et al. Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography. Sci Rep 2020 Nov 24;10(1):20495 [FREE Full text] [doi: 10.1038/s41598-020-77599-6] [Medline: 33235279]
151. Makimoto H, Höckmann M, Lin T, Glöckner D, Gerguri S, Clasen L, et al. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. Sci Rep 2020 May 21;10(1):8445 [FREE Full text] [doi: 10.1038/s41598-020-65105-x] [Medline: 32439873]
152. Fu L, Lu B, Nie B, Peng Z, Liu H, Pi X. Hybrid network with attention mechanism for detection and location of myocardial infarction based on 12-lead electrocardiogram signals. Sensors (Basel) 2020 Feb 14;20(4):1020 [FREE Full text] [doi: 10.3390/s20041020] [Medline: 32074979]
153. Raghunath S, Ulloa Cerna AE, Jing L, vanMaanen DP, Stough J, Hartzel DN, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nat Med 2020 Jun;26(6):886-891 [FREE Full text] [doi: 10.1038/s41591-020-0870-z] [Medline: 32393799]
154. Kwon JM, Kim KH, Akkus Z, Jeon KH, Park J, Oh BH. Artificial intelligence for detecting mitral regurgitation using electrocardiography. J Electrocardiol 2020;59:151-157 [FREE Full text] [doi: 10.1016/j.jelectrocard.2020.02.008] [Medline: 32146201]
155. Missel R, Gyawali PK, Murkute JV, Li Z, Zhou S, AbdelWahab A, et al. A hybrid machine learning approach to localizing the origin of ventricular tachycardia using 12-lead electrocardiograms. Comput Biol Med 2020 Nov;126:104013 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.104013] [Medline: 33002841]
156. Çınar A, Tuncer SA. Classification of normal sinus rhythm, abnormal arrhythmia and congestive heart failure ECG signals using LSTM and hybrid CNN-SVM deep neural networks. Comput Methods Biomech Biomed Engin 2021 Feb;24(2):203-214 [FREE Full text] [doi: 10.1080/10255842.2020.1821192] [Medline: 32955928]
157. Cho J, Lee B, Kwon JM, Lee Y, Park H, Oh BH, et al. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. ASAIO J 2021 Mar 01;67(3):314-321 [FREE Full text] [doi: 10.1097/MAT.0000000000001218] [Medline: 33627606]
158. Gumpfer N, Grün D, Hannig J, Keller T, Guckert M. Detecting myocardial scar using electrocardiogram data and deep neural networks. Biol Chem 2021 Jul 27;402(8):911-923 [FREE Full text] [doi: 10.1515/hsz-2020-0169] [Medline: 33006947]
159. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, et al. Assessing and mitigating bias in medical artificial intelligence. Circ Arrhythmia Electrophysiology 2020 Mar;13(3):e007988 [FREE Full text] [doi: 10.1161/circep.119.007988]
160. Han C, Song Y, Lim H, Tae Y, Jang J, Lee BT, et al. Automated detection of acute myocardial infarction using asynchronous electrocardiogram signals-preview of implementing artificial intelligence with multichannel electrocardiographs obtained

from smartwatches: retrospective study. J Med Internet Res 2021 Sep 10;23(9):e31129 [FREE Full text] [doi: 10.2196/31129] [Medline: 34505839]
161. Ivaturi P, Gadaleta M, Pandey A, Pazzani M, Steinhubl S, Quer G. A comprehensive explanation framework for biomedical time series classification. IEEE J Biomed Health Inform 2021 Jul;25(7):2398-2408 [FREE Full text] [doi: 10.1109/jbhi.2021.3060997]
162. Baek YS, Lee SC, Choi W, Kim DH. A new deep learning algorithm of 12-lead electrocardiogram for identifying atrial fibrillation during sinus rhythm. Sci Rep 2021 Jun 17;11(1):12818 [FREE Full text] [doi: 10.1038/s41598-021-92172-5] [Medline: 34140578]
163. Radhakrishnan T, Karhade J, Ghosh S, Muduli P, Tripathy R, Acharya UR. AFCNNet: automated detection of AF using chirplet transform and deep convolutional bidirectional long short term memory network with ECG signals. Comput Biol Med 2021 Oct;137:104783 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104783] [Medline: 34481184]
164. Tutuko B, Nurmaini S, Tondas AE, Rachmatullah MN, Darmawahyuni A, Esafri R, et al. AFibNet: an implementation of atrial fibrillation detection with convolutional neural network. BMC Med Inform Decis Mak 2021 Jul 14;21(1):216 [FREE Full text] [doi: 10.1186/s12911-021-01571-1] [Medline: 34261486]
165. Salinas-Martínez R, de Bie J, Marzocchi N, Sandberg F. Detection of brief episodes of atrial fibrillation based on electrocardiomatrix and convolutional neural network. Front Physiol 2021;12:673819 [FREE Full text] [doi: 10.3389/fphys.2021.673819] [Medline: 34512372]
166. Seo HC, Oh S, Kim H, Joo S. ECG data dependency for atrial fibrillation detection based on residual networks. Sci Rep 2021 Sep 14;11(1):18256 [FREE Full text] [doi: 10.1038/s41598-021-97308-1] [Medline: 34521892]
167. Jo YY, Cho Y, Lee SY, Kwon JM, Kim KH, Jeon KH, et al. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. Int J Cardiol 2021 Apr 01;328:104-110 [FREE Full text] [doi: 10.1016/j.ijcard.2020.11.053] [Medline: 33271204]
168. Zhang X, Li J, Cai Z, Zhang L, Chen Z, Liu C. Over-fitting suppression training strategies for deep learning-based atrial fibrillation detection. Med Biol Eng Comput 2021 Jan;59(1):165-173 [FREE Full text] [doi: 10.1007/s11517-020-02292-9] [Medline: 33387183]
169. Zhang H, Dong Z, Sun M, Gu H, Wang Z. TP-CNN: a detection method for atrial fibrillation based on transposed projection signals with compressed sensed ECG. Comput Methods Programs Biomed 2021 Oct;210:106358 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106358] [Medline: 34478912]
170. Petmezas G, Haris K, Stefanopoulos L, Kilintzis V, Tzavelis A, Rogers J, et al. Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets. Biomedical Signal Process Control 2021 Jan;63:102194 [FREE Full text] [doi: 10.1016/j.bspc.2020.102194]
171. Nishimori M, Kiuchi K, Nishimura K, Kusano K, Yoshida A, Adachi K, et al. Accessory pathway analysis using a multimodal deep learning model. Sci Rep 2021 Apr 13;11(1):8045 [FREE Full text] [doi: 10.1038/s41598-021-87631-y] [Medline: 33850245]
172. Sawano S, Kodera S, Katsushika S, Nakamoto M, Ninomiya K, Shinohara H, et al. Deep learning model to detect significant aortic regurgitation using electrocardiography. J Cardiol 2022 Mar;79(3):334-341 [FREE Full text] [doi: 10.1016/j.jjcc.2021.08.029] [Medline: 34544652]
173. Yang X, Zhang X, Yang M, Zhang L. 12-Lead ECG arrhythmia classification using cascaded convolutional neural network and expert feature. J Electrocardiol 2021;67:56-62 [FREE Full text] [doi: 10.1016/j.jelectrocard.2021.04.016] [Medline: 34082153]
174. Kiyasseh D, Zhu T, Clifton D. A clinical deep learning framework for continually learning from cardiac signals across diseases, time, modalities, and institutions. Nat Commun 2021 Jul 09;12(1):4221 [FREE Full text] [doi: 10.1038/s41467-021-24483-0] [Medline: 34244504]
175. Che C, Zhang P, Zhu M, Qu Y, Jin B. Constrained transformer network for ECG signal processing and arrhythmia classification. BMC Med Inform Decis Mak 2021 Jun 09;21(1):184 [FREE Full text] [doi: 10.1186/s12911-021-01546-2] [Medline: 34107920]
176. Jo YY, Kwon JM, Jeon KH, Cho YH, Shin JH, Lee YJ, et al. Detection and classification of arrhythmia using an explainable deep learning model. J Electrocardiol 2021;67:124-132 [FREE Full text] [doi: 10.1016/j.jelectrocard.2021.06.006] [Medline: 34225095]
177. Mousavi S, Afghah F, Khadem F, Acharya UR. ECG Language processing (ELP): a new technique to analyze ECG signals. Comput Methods Programs Biomed 2021 Apr;202:105959 [FREE Full text] [doi: 10.1016/j.cmpb.2021.105959] [Medline: 33607552]
178. Jang JH, Kim TY, Yoon D. Effectiveness of transfer learning for deep learning-based electrocardiogram analysis. Healthc Inform Res 2021 Jan;27(1):19-28 [FREE Full text] [doi: 10.4258/hir.2021.27.1.19] [Medline: 33611873]
179. Jiang M, Gu J, Li Y, Wei B, Zhang J, Wang Z, et al. HADLN: hybrid attention-based deep learning network for automated arrhythmia classification. Front Physiol 2021 Jul 5;12:683025 [FREE Full text] [doi: 10.3389/fphys.2021.683025] [Medline: 34290619]

180. Lu P, Gao Y, Xi H, Zhang Y, Gao C, Zhou B, et al. KecNet: a light neural network for arrhythmia classification based on knowledge reinforcement. J Healthc Eng 2021 Apr 24;2021:6684954-6684910 [FREE Full text] [doi: 10.1155/2021/6684954] [Medline: 33995984]
181. Lee H, Shin M. Learning explainable time-morphology patterns for automatic arrhythmia classification from short single-lead ECGs. Sensors (Basel) 2021 Jun 24;21(13):4331 [FREE Full text] [doi: 10.3390/s21134331] [Medline: 34202805]
182. Zhang J, Liang D, Liu A, Gao M, Chen X, Zhang X, et al. MLBF-Net: a multi-lead-branch fusion network for multi-class arrhythmia classification using 12-lead ECG. IEEE J Transl Eng Health Med 2021;9:1-11 [FREE Full text] [doi: 10.1109/jtehm.2021.3064675]
183. Luo X, Yang L, Cai H, Tang R, Chen Y, Li W. Multi-classification of arrhythmias using a HCRNet on imbalanced ECG datasets. Comput Methods Programs Biomed 2021 Sep;208:106258 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106258] [Medline: 34218172]
184. Zhang H, Liu C, Zhang Z, Xing Y, Liu X, Dong R, et al. Recurrence plot-based approach for cardiac arrhythmia classification using inception-ResNet-v2. Front Physiol 2021;12:648950 [FREE Full text] [doi: 10.3389/fphys.2021.648950] [Medline: 34079470]
185. Wang J, Li R, Li R, Fu B, Xiao C, Chen D. Towards interpretable arrhythmia classification with human-machine collaborative knowledge representation. IEEE Trans Biomed Eng 2021 Jul;68(7):2098-2109 [FREE Full text] [doi: 10.1109/tbme.2020.3024970]
186. Chang KC, Hsieh PH, Wu MY, Wang YC, Chen JY, Tsai FJ, et al. Usefulness of machine learning-based detection and classification of cardiac arrhythmias with 12-lead electrocardiograms. Can J Cardiol 2021 Jan;37(1):94-104 [FREE Full text] [doi: 10.1016/j.cjca.2020.02.096] [Medline: 32585216]
187. Elul Y, Rosenberg AA, Schuster A, Bronstein AM, Yaniv Y. Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning-based ECG analysis. Proc Natl Acad Sci U S A 2021 Jun 15;118(24):e2020620118 [FREE Full text] [doi: 10.1073/pnas.2020620118] [Medline: 34099565]
188. Nannavecchia A, Girardi F, Fina PR, Scalera M, Dimauro G. Personal heart health monitoring based on 1D convolutional neural network. J Imaging 2021 Feb 05;7(2):26 [FREE Full text] [doi: 10.3390/jimaging7020026] [Medline: 34460625]
189. Yoo J, Jun TJ, Kim Y. xECGNet: fine-tuning attention map within convolutional neural network to improve detection and explainability of concurrent cardiac arrhythmias. Comput Methods Programs Biomed 2021 Sep;208:106281. [doi: 10.1016/j.cmpb.2021.106281] [Medline: 34333207]
190. Mori H, Inai K, Sugiyama H, Muragaki Y. Diagnosing atrial septal defect from electrocardiogram with deep learning. Pediatr Cardiol 2021 Aug;42(6):1379-1387 [FREE Full text] [doi: 10.1007/s00246-021-02622-0] [Medline: 33907875]
191. Liu C, Liu C, Hu K, Tseng VS, Chang S, Lin Y, et al. A deep learning-enabled electrocardiogram model for the identification of a rare inherited arrhythmia: Brugada syndrome. Can J Cardiol 2022 Feb;38(2):152-159 [FREE Full text] [doi: 10.1016/j.cjca.2021.08.014] [Medline: 34461230]
192. Jahmunah V, Ng EY, San TR, Acharya UR. Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals. Comput Biol Med 2021 Jul;134:104457. [doi: 10.1016/j.compbiomed.2021.104457] [Medline: 33991857]
193. Bender T, Seidler T, Bengel P, Sax U, Krefting D. Application of pre-trained deep learning models for clinical ECGs. Stud Health Technol Inform 2021 Sep 21;283:39-45 [FREE Full text] [doi: 10.3233/SHTI210539] [Medline: 34545818]
194. Fu Z, Hong S, Zhang R, Du S. Artificial-intelligence-enhanced mobile system for cardiovascular health management. Sensors (Basel) 2021 Jan 24;21(3):773 [FREE Full text] [doi: 10.3390/s21030773] [Medline: 33498892]
195. Dai H, Hwang HG, Tseng VS. Convolutional neural network based automatic screening tool for cardiovascular diseases using different intervals of ECG signals. Comput Methods Programs Biomed 2021 May;203:106035 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106035] [Medline: 33770545]
196. Deevi SA, Kaniraja CP, Mani VD, Mishra D, Ummar S, Satheesh C. HeartNetEC: a deep representation learning approach for ECG beat classification. Biomed Eng Lett 2021 Feb 08;11(1):69-84 [FREE Full text] [doi: 10.1007/s13534-021-00184-x] [Medline: 33747604]
197. Chen CY, Lin YT, Lee SJ, Tsai WC, Huang TC, Liu YH, et al. Automated ECG classification based on 1D deep learning network. Methods 2022 Jun;202:127-135 [FREE Full text] [doi: 10.1016/j.ymeth.2021.04.021] [Medline: 33930574]
198. Wang J, Qiao X, Liu C, Wang X, Liu Y, Yao L, et al. Automated ECG classification using a non-local convolutional block attention module. Comput Methods Programs Biomed 2021 May;203:106006. [doi: 10.1016/j.cmpb.2021.106006] [Medline: 33735660]
199. Wang T, Lu C, Sun Y, Yang M, Liu C, Ou C. Automatic ECG classification using continuous wavelet transform and convolutional neural network. Entropy (Basel) 2021 Jan 18;23(1):119 [FREE Full text] [doi: 10.3390/e23010119] [Medline: 33477566]
200. Pokaprakarn T, Kitzmiller RR, Moorman JR, Lake DE, Krishnamurthy AK, Kosorok MR. Sequence to sequence ECG cardiac rhythm classification using convolutional recurrent neural networks. IEEE J Biomed Health Inform 2022 Feb;26(2):572-580 [FREE Full text] [doi: 10.1109/jbhi.2021.3098662]
201. Weimann K, Conrad TO. Transfer learning for ECG classification. Sci Rep 2021 Mar 04;11(1):5251 [FREE Full text] [doi: 10.1038/s41598-021-84374-8] [Medline: 33664343]

202. Zhang D, Yang S, Yuan X, Zhang P. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. iScience 2021 Apr 23;24(4):102373 [FREE Full text] [doi: 10.1016/j.isci.2021.102373] [Medline: 33981967]
203. Mishra S, Khatwani G, Patil R, Sapariya D, Shah V, Parmar D, et al. ECG paper record digitization and diagnosis using deep learning. J Med Biol Eng 2021;41(4):422-432 [FREE Full text] [doi: 10.1007/s40846-021-00632-0] [Medline: 34149335]
204. van de Leur RR, Taha K, Bos MN, van der Heijden JF, Gupta D, Cramer MJ, et al. Discovering and visualizing disease-specific electrocardiogram features using deep learning. Circ Arrhythmia Electrophysiol 2021 Feb;14(2):e009056 [FREE Full text] [doi: 10.1161/circep.120.009056]
205. Zhang J, Liu A, Liang D, Chen X, Gao M. Interpatient ECG heartbeat classification with an adversarial convolutional neural network. J Healthc Eng 2021;2021:9946596 [FREE Full text] [doi: 10.1155/2021/9946596] [Medline: 34194685]
206. Ammour N, Alhichri H, Bazi Y, Alajlan N. LwF-ECG: learning-without-forgetting approach for electrocardiogram heartbeat classification based on memory with task selector. Comput Biol Med 2021 Oct;137:104807 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104807] [Medline: 34496312]
207. Wu M, Lu Y, Yang W, Wong SY. A study on arrhythmia via ECG signal classification using the convolutional neural network. Front Comput Neurosci 2020;14:564015 [FREE Full text] [doi: 10.3389/fncom.2020.564015] [Medline: 33469423]
208. Ma H, Chen C, Zhu Q, Yuan H, Chen L, Shu M. An ECG signal classification method based on dilated causal convolution. Comput Math Methods Med 2021;2021:6627939 [FREE Full text] [doi: 10.1155/2021/6627939] [Medline: 33603825]
209. Siontis KC, Liu K, Bos JM, Attia ZI, Cohen-Shelly M, Arruda-Olson AM, et al. Detection of hypertrophic cardiomyopathy by an artificial intelligence electrocardiogram in children and adolescents. Int J Cardiol 2021 Oct 01;340:42-47 [FREE Full text] [doi: 10.1016/j.ijcard.2021.08.026] [Medline: 34419527]
210. Huang Y, Li H, Yu X. A multiview feature fusion model for heartbeat classification. Physiol Meas 2021 Jun 29;42(6). [doi: 10.1088/1361-6579/ac010f] [Medline: 33984841]
211. Zhang Y, Li J, Wei S, Zhou F, Li D. Heartbeats classification using hybrid time-frequency analysis and transfer learning based on ResNet. IEEE J Biomed Health Inform 2021 Nov;25(11):4175-4184 [FREE Full text] [doi: 10.1109/jbhi.2021.3085318]
212. Wang CX, Zhang YC, Kong QL, Wu ZL, Yang PP, Zhu CH, et al. Development and validation of a deep learning model to screen hypokalemia from electrocardiogram in emergency patients. Chin Med J (Engl) 2021 Sep 02;134(19):2333-2339 [FREE Full text] [doi: 10.1097/CM9.0000000000001650] [Medline: 34483253]
213. Paragliola G, Coronato A. An hybrid ECG-based deep network for the early identification of high-risk to major cardiovascular events for hypertension patients. J Biomed Inform 2021 Jan;113:103648 [FREE Full text] [doi: 10.1016/j.jbi.2020.103648] [Medline: 33276113]
214. Sun JY, Qiu Y, Guo HC, Hua Y, Shao B, Qiao YC, et al. A method to screen left ventricular dysfunction through ECG based on convolutional neural network. J Cardiovasc Electrophysiol 2021 Apr;32(4):1095-1102 [FREE Full text] [doi: 10.1111/jce.14936] [Medline: 33565217]
215. Attia IZ, Tseng AS, Benavente ED, Medina-Inojosa JR, Clark TG, Malyutina S, et al. External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. Int J Cardiol 2021 Apr 15;329:130-135 [FREE Full text] [doi: 10.1016/j.ijcard.2020.12.065] [Medline: 33400971]
216. Bigler MR, Seiler C. Detection of myocardial ischemia by intracoronary ECG using convolutional neural networks. Eur Heart J 2021;42(Supplement_1) [FREE Full text] [doi: 10.1093/eurheartj/ehab724.3049]
217. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, Nemani A, Carbonati T, Jing L, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation–related stroke. Circulation 2021 Mar 30;143(13):1287-1298 [FREE Full text] [doi: 10.1161/circulationaha.120.047829]
218. Yang J, Cai W, Wang M. Premature beats detection based on a novel convolutional neural network. Physiol Meas 2021 Jul 28;42(7) [FREE Full text] [doi: 10.1088/1361-6579/ac0e82] [Medline: 34167103]
219. Yu J, Wang X, Chen X, Guo J. Automatic premature ventricular contraction detection using deep metric learning and KNN. Biosensors (Basel) 2021 Mar 04;11(3):69 [FREE Full text] [doi: 10.3390/bios11030069] [Medline: 33806367]
220. Naz M, Shah JH, Khan MA, Sharif M, Raza M, Damaševičius R. From ECG signals to images: a transformation based approach for deep learning. PeerJ Comput Sci 2021;7:e386 [FREE Full text] [doi: 10.7717/peerj-cs.386] [Medline: 33817032]
221. Petryshak B, Kachko I, Maksymenko M, Dobosevych O. Robust deep learning pipeline for PVC beats localization. Technol Health Care 2021 Mar 25;29:475-486 [FREE Full text] [doi: 10.3233/thc-218045]
222. Sabut S, Pandey O, Mishra BS, Mohanty M. Detection of ventricular arrhythmia using hybrid time-frequency-based features and deep neural network. Phys Eng Sci Med 2021 Mar;44(1):135-145 [FREE Full text] [doi: 10.1007/s13246-020-00964-2] [Medline: 33417159]
223. Liu WT, Lin CS, Tsao TP, Lee CC, Cheng CC, Chen JT, et al. A deep-learning algorithm-enhanced system integrating electrocardiograms and chest X-rays for diagnosing aortic dissection. Can J Cardiol 2022 Feb;38(2):160-168 [FREE Full text] [doi: 10.1016/j.cjca.2021.09.028] [Medline: 34619339]
224. Li Z, Wang H, Liu X. A one-dimensional Siamese few-shot learning approach for ECG classification under limited data. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology

Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630622 [doi: 10.1109/embc46164.2021.9630622]
225. Liu X, Wang H, Li Z. An approach for deep learning in ECG classification tasks in the presence of noisy labels. In: Proceedngs of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630763 [doi: 10.1109/embc46164.2021.9630763]
226. Liu WC, Lin C, Lin CS, Tsai MC, Chen SJ, Tsai SH, et al. An artificial intelligence-based alarm strategy facilitates management of acute myocardial infarction. J Pers Med 2021 Nov 04;11(11):1149 [FREE Full text] [doi: 10.3390/jpm11111149] [Medline: 34834501]
227. Krasteva V, Christov I, Naydenov S, Stoyanov T, Jekova I. Application of dense neural networks for detection of atrial fibrillation and ranking of augmented ECG feature set. Sensors (Basel) 2021 Oct 15;21(20):6848 [FREE Full text] [doi: 10.3390/s21206848] [Medline: 34696061]
228. Ramesh J, Solatidehkordi Z, Aburukba R, Sagahyroon A. Atrial fibrillation classification with smart wearables using short-term heart rate variability and deep convolutional neural networks. Sensors (Basel) 2021 Oct 30;21(21):7233 [FREE Full text] [doi: 10.3390/s21217233] [Medline: 34770543]
229. Xie Y, Qin L, Tan H, Li X, Liu B, Wang H. Automatic 12-leading electrocardiogram classification network with deformable convolution. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630227 [doi: 10.1109/embc46164.2021.9630227]
230. Liu Y, Li Q, Wang K, Liu J, He R, Yuan Y, et al. Automatic multi-label ECG classification with category imbalance and cost-sensitive thresholding. Biosensors (Basel) 2021 Nov 14;11(11):453 [FREE Full text] [doi: 10.3390/bios11110453] [Medline: 34821669]
231. Ullah W, Siddique I, Zulqarnain RM, Alam MM, Ahmad I, Raza UA. Classification of arrhythmia in heartbeat detection using deep learning. Comput Intell Neurosci 2021;2021:2195922 [FREE Full text] [doi: 10.1155/2021/2195922] [Medline: 34712316]
232. Tadesse GA, Javed H, Weldemariam K, Liu Y, Liu J, Chen J, et al. DeepMI: deep multi-lead ECG fusion for identifying myocardial infarction and its occurrence-time. Artif Intell Med 2021 Nov;121:102192 [FREE Full text] [doi: 10.1016/j.artmed.2021.102192] [Medline: 34763807]
233. Adedinsewo DA, Johnson PW, Douglass EJ, Attia IZ, Phillips SD, Goswami RM, et al. Detecting cardiomyopathies in pregnancy and the postpartum period with an electrocardiogram-based deep learning model. Eur Heart J Digit Health 2021 Dec;2(4):586-596 [FREE Full text] [doi: 10.1093/ehjdh/ztab078] [Medline: 34993486]
234. Chen L, Yu H, Huang Y, Jin H. ECG signal-enabled automatic diagnosis technology of heart failure. J Healthc Eng 2021;2021:5802722 [FREE Full text] [doi: 10.1155/2021/5802722] [Medline: 34777736]
235. Akbilgic O, Butler L, Karabayir I, Chang PP, Kitzman DW, Alonso A, et al. ECG-AI: electrocardiographic artificial intelligence model for prediction of heart failure. Eur Heart J Digit Health 2021;2(4):626-634 [FREE Full text] [doi: 10.1093/ehjdh/ztab080]
236. Khurshid S, Friedman S, Reeder C, Di Achille P, Diamant N, Singh P, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. Circulation 2022 Jan 11;145(2):122-133 [FREE Full text] [doi: 10.1161/circulationaha.121.057480]
237. Gibson CM, Mehta S, Ceschim MR, Frauenfelder A, Vieira D, Botelho R, et al. Evolution of single-lead ECG for STEMI detection using a deep learning approach. Int J Cardiol 2022 Jan 01;346:47-52 [FREE Full text] [doi: 10.1016/j.ijcard.2021.11.039] [Medline: 34801613]
238. Zhang P, Ma C, Sun Y, Fan G, Song F, Feng Y, et al. Global hybrid multi-scale convolutional network for accurate and robust detection of atrial fibrillation using single-lead ECG recordings. Comput Biol Med 2021 Dec;139:104880 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104880] [Medline: 34700255]
239. Bizzego A, Gabrieli G, Neoh MJ, Esposito G. Improving the efficacy of deep-learning models for heart beat detection on heterogeneous datasets. Bioengineering (Basel) 2021 Nov 28;8(12):193 [FREE Full text] [doi: 10.3390/bioengineering8120193] [Medline: 34940346]
240. Li H, Wang X, Liu C, Li P, Jiao Y. Integrating multi-domain deep features of electrocardiogram and phonocardiogram for coronary artery disease detection. Comput Biol Med 2021 Nov;138:104914 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104914] [Medline: 34638021]
241. Li Y, Qian R, Li K. Inter-patient arrhythmia classification with improved deep residual convolutional neural network. Comput Methods Programs Biomed 2022 Feb;214:106582 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106582] [Medline: 34933228]
242. Lai C, Zhou S, Trayanova NA. Optimal ECG-lead selection increases generalizability of deep learning on ECG abnormality classification. Philos Trans A Math Phys Eng Sci 2021 Dec 13;379(2212):20200258 [FREE Full text] [doi: 10.1098/rsta.2020.0258] [Medline: 34689629]

243. Tzou HA, Lin SF, Chen PS. Paroxysmal atrial fibrillation prediction based on morphological variant P-wave analysis with wideband ECG and deep learning. Comput Methods Programs Biomed 2021 Nov;211:106396 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106396] [Medline: 34592687]
244. Bollepalli SC, Sevakula RK, Au - Yeung WM, Kassab MB, Merchant FM, Bazoukis G, et al. Real - time arrhythmia detection using hybrid convolutional neural networks. J Am Heart Assoc 2021 Dec 07;10(23):e023222 [FREE Full text] [doi: 10.1161/jaha.121.023222]
245. Malik J, Devecioglu OC, Kiranyaz S, Ince T, Gabbouj M. Real-time patient-specific ECG classification by 1d self-operational neural networks. IEEE Trans Biomed Eng 2022 May;69(5):1788-1801 [FREE Full text] [doi: 10.1109/tbme.2021.3135622]
246. Luo C, Wang G, Ding Z, Chen H, Yang F. Segment origin prediction: a self-supervised learning method for electrocardiogram arrhythmia classification. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630616 [doi: 10.1109/embc46164.2021.9630616]
247. Lee BT, Kong ST, Song Y, Lee Y. Self-supervised learning with electrocardiogram delineation for arrhythmia detection. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630364 [doi: 10.1109/embc46164.2021.9630364]
248. Rasmussen SM, Jensen M, Meyhoff CS, Aasvang EK, Słrensen H. Semi-supervised analysis of the electrocardiogram using deep generative models. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9629915 [doi: 10.1109/embc46164.2021.9629915]
249. Park J, An J, Kim J, Jung S, Gil Y, Jang Y, et al. Study on the use of standard 12-lead ECG data for rhythm-type ECG classification problems. Comput Methods Programs Biomed 2022 Feb;214:106521 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106521] [Medline: 34844765]
250. Vaid A, Johnson KW, Badgeley MA, Somani SS, Bicak M, Landi I, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. JACC Cardiovasc Imaging 2022 Mar;15(3):395-410 [FREE Full text] [doi: 10.1016/j.jcmg.2021.08.004] [Medline: 34656465]
251. Teplitzky BA, McRoberts M, Ghanbari H. Deep learning for comprehensive ECG annotation. Heart Rhythm 2020 May;17(5 Pt B):881-888 [FREE Full text] [doi: 10.1016/j.hrthm.2020.02.015] [Medline: 32354454]
252. Li Y, Qu Q, Wang M, Yu L, Wang J, Shen L, et al. Deep learning for digitizing highly noisy paper-based ECG records. Comput Biol Med 2020 Dec;127:104077 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.104077] [Medline: 33171291]
253. Cao F, Budhota A, Chen H, Singh Rajput K. Feature matching based ECG generative network for arrhythmia event augmentation. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 20-24, 2020; Montreal, QC, Canada URL: https://doi.org/10.1109/embc44109.2020.9175668 [doi: 10.1109/embc44109.2020.9175668]
254. Herraiz AH, Martínez-Rodrigo A, Bertomeu-González V, Quesada A, Rieta JJ, Alcaraz R. A deep learning approach for featureless robust quality assessment of intermittent atrial fibrillation recordings from portable and wearable devices. Entropy (Basel) 2020 Jul 01;22(7):733 [FREE Full text] [doi: 10.3390/e22070733] [Medline: 33286505]
255. Fotiadou E, Konopczyński T, Hesser J, Vullings R. End-to-end trained encoder-decoder convolutional neural network for fetal electrocardiogram signal denoising. Physiol Meas 2020 Feb 05;41(1):015005 [FREE Full text] [doi: 10.1088/1361-6579/ab69b9] [Medline: 31918422]
256. Fotiadou E, Vullings R. Multi-channel fetal ECG denoising with deep convolutional neural networks. Front Pediatr 2020;8:508 [FREE Full text] [doi: 10.3389/fped.2020.00508] [Medline: 32984218]
257. Vo K, Le T, Rahmani AM, Dutt N, Cao H. An efficient and robust deep learning method with 1-D octave convolution to extract fetal electrocardiogram. Sensors (Basel) 2020 Jul 04;20(13):3757 [FREE Full text] [doi: 10.3390/s20133757] [Medline: 32635568]
258. Murat F, Yildirim O, Talo M, Baloglu UB, Demir Y, Acharya UR. Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. Comput Biol Med 2020 May;120:103726 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103726] [Medline: 32421643]
259. Silva P, Luz E, Silva G, Moreira G, Wanner E, Vidal F, et al. Towards better heartbeat segmentation with deep learning classification. Sci Rep 2020 Nov 26;10(1):20701 [FREE Full text] [doi: 10.1038/s41598-020-77745-0] [Medline: 33244078]
260. Hao C, Wibowo S, Singh rajput K. Compressive sampling based multi-spectrum deep learning for sub-nyquist pacemaker ECG analysis. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 20-24, 2020; Montreal, QC, Canada URL: https://doi.org/10.1109/embc44109.2020.9175625 [doi: 10.1109/embc44109.2020.9175625]

261. Vijayarangan S, Vignesh R, Murugesan B, Preejith SP, Joseph J, Sivaprakasam M. RPnet: a deep learning approach for robust R peak detection in noisy ECG. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 20-24, 2020; Montreal, QC, Canada URL: https://doi.org/10.1109/embc44109.2020.9176084 [doi: 10.1109/embc44109.2020.9176084]
262. Zaman SD, Morshed BI. Estimating reliability of signal quality of physiological data from data statistics itself for real-time wearables. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 20-24, 2020; Montreal, QC, Canada URL: https://doi.org/10.1109/embc44109.2020.9175317 [doi: 10.1109/embc44109.2020.9175317]
263. Hicks SA, Isaksen JL, Thambawita V, Ghouse J, Ahlberg G, Linneberg A, et al. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. medRxiv 2021 [FREE Full text] [doi: 10.1101/2021.01.06.20248927]
264. Gyawali PK, Murkute JV, Toloubidokhti M, Jiang X, Horacek BM, Sapp JL, et al. Learning to disentangle inter-subject anatomical variations in electrocardiographic data. IEEE Trans Biomed Eng 2022 Feb;69(2):860-870 [FREE Full text] [doi: 10.1109/tbme.2021.3108164]
265. Jimenez-Perez G, Alcaine A, Camara O. Delineation of the electrocardiogram with a mixed-quality-annotations dataset using convolutional neural networks. Sci Rep 2021 Jan 13;11(1):863 [FREE Full text] [doi: 10.1038/s41598-020-79512-7] [Medline: 33441632]
266. Kuznetsov VV, Moskalenko VA, Gribanov DV, Zolotykh NY. Interpretable feature generation in ECG using a variational autoencoder. Front Genet 2021;12:638191 [FREE Full text] [doi: 10.3389/fgene.2021.638191] [Medline: 33868375]
267. Liu G, Han X, Tian L, Zhou W, Liu H. ECG quality assessment based on hand-crafted statistics and deep-learned S-transform spectrogram features. Comput Methods Programs Biomed 2021 Sep;208:106269 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106269] [Medline: 34298474]
268. Seeuws N, De Vos M, Bertrand A. Electrocardiogram quality assessment using unsupervised deep learning. IEEE Trans Biomed Eng 2022 Feb;69(2):882-893 [FREE Full text] [doi: 10.1109/tbme.2021.3108621]
269. Bacoyannis T, Ly B, Cedilnik N, Cochet H, Sermesant M. Deep learning formulation of electrocardiographic imaging integrating image and signal information with data-driven regularization. Europace 2021 Mar 04;23(23 Suppl 1):i55-i62 [FREE Full text] [doi: 10.1093/europace/euaa391] [Medline: 33751073]
270. Rjoob K, Bond R, Finlay D, McGilligan V, J Leslie S, Rababah A, et al. Reliable deep learning-based detection of misplaced chest electrodes during electrocardiogram recording: algorithm development and validation. JMIR Med Inform 2021 Apr 16;9(4):e25347 [FREE Full text] [doi: 10.2196/25347] [Medline: 33861205]
271. Fotiadou E, van Sloun RJ, van Laar JO, Vullings R. A dilated inception CNN-LSTM network for fetal heart rate estimation. Physiol Meas 2021 May 13;42(4):045007 [FREE Full text] [doi: 10.1088/1361-6579/abf7db] [Medline: 33853039]
272. Giudicessi JR, Schram M, Bos JM, Galloway CD, Shreibati JB, Johnson PW, et al. Artificial intelligence–enabled assessment of the heart rate corrected QT interval using a mobile electrocardiogram device. Circulation 2021 Mar 30;143(13):1274-1286 [FREE Full text] [doi: 10.1161/circulationaha.120.050231]
273. Ganapathy N, Swaminathan R, Deserno TM. Adaptive learning and cross training improves R-wave detection in ECG. Comput Methods Programs Biomed 2021 Mar;200:105931 [FREE Full text] [doi: 10.1016/j.cmpb.2021.105931] [Medline: 33508772]
274. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. IEEE J Biomed Health Inform 2021 May;25(5):1519-1528 [FREE Full text] [doi: 10.1109/jbhi.2020.3022989]
275. Śmigiel S, Pałczyński K, Ledziński D. Deep learning techniques in the classification of ECG signals using r-peak detection based on the PTB-XL dataset. Sensors (Basel) 2021 Dec 07;21(24):8174 [FREE Full text] [doi: 10.3390/s21248174] [Medline: 34960267]
276. Pool MD, de Vos BD, Winter MM, Isgum I. Deep learning-based data-point precise R-peak detection in single-lead electrocardiograms. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630062 [doi: 10.1109/embc46164.2021.9630062]
277. Spicher N, Klingenberg A, Purrucker V, Deserno TM. Edge computing in 5G cellular networks for real-time analysis of electrocardiography recorded with wearable textile sensors. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630875 [doi: 10.1109/embc46164.2021.9630875]
278. Venton J, Harris PM, Sundar A, Smith NA, Aston PJ. Robustness of convolutional neural networks to physiological electrocardiogram noise. Philos Trans A Math Phys Eng Sci 2021 Dec 13;379(2212):20200262 [FREE Full text] [doi: 10.1098/rsta.2020.0262] [Medline: 34689617]
279. Mehari T, Strodthoff N. Self-supervised representation learning from 12-lead ECG data. Comput Biol Med 2022 Feb;141:105114 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.105114] [Medline: 34973584]

280. M Jomaa R, Mathkour H, Bazi Y, Islam MS. End-to-end deep learning fusion of fingerprint and electrocardiogram signals for presentation attack detection. Sensors (Basel) 2020 Apr 07;20(7):2085 [FREE Full text] [doi: 10.3390/s20072085] [Medline: 32272813]
281. Song HK, AlAlkeem E, Yun J, Kim TH, Yoo H, Heo D, et al. Deep user identification model with multiple biometric data. BMC Bioinformatics 2020 Jul 16;21(1):315 [FREE Full text] [doi: 10.1186/s12859-020-03613-3] [Medline: 32677882]
282. Belo D, Bento N, Silva H, Fred A, Gamboa H. ECG biometrics using deep learning and relative score threshold classification. Sensors (Basel) 2020 Jul 22;20(15):4078 [FREE Full text] [doi: 10.3390/s20154078] [Medline: 32707861]
283. AlDuwaile DA, Islam MS. Using convolutional neural network and a single heartbeat for ECG biometric recognition. Entropy (Basel) 2021 Jun 09;23(6):733 [FREE Full text] [doi: 10.3390/e23060733] [Medline: 34207846]
284. Wu S, Wei S, Chang C, Swindlehurst AL, Chiu J. A scalable open-set ECG identification system based on compressed CNNs. IEEE Trans Neural Netw Learning Syst 2021:1-15 [FREE Full text] [doi: 10.1109/tnnls.2021.3127497]
285. Chiu JK, Chang CS, Wu SC. ECG-based biometric recognition without QRS segmentation: a deep learning-based approach. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630899 [doi: 10.1109/embc46164.2021.9630899]
286. Ghazarian A, Zheng J, El-Askary H, Chu H, Fu G, Rakovski C. Increased risks of re-identification for patients posed by deep learning-based ECG identification algorithms. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630880 [doi: 10.1109/embc46164.2021.9630880]
287. Fonseca P, van Gilst MM, Radha M, Ross M, Moreau A, Cerny A, et al. Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population. Sleep 2020 Sep 14;43(9):zsaa048 [FREE Full text] [doi: 10.1093/sleep/zsaa048] [Medline: 32249911]
288. Sridhar N, Shoeb A, Stephens P, Kharbouch A, Shimol DB, Burkart J, et al. Deep learning for automated sleep staging using instantaneous heart rate. NPJ Digit Med 2020 Aug 20;3(1):106 [FREE Full text] [doi: 10.1038/s41746-020-0291-x]
289. Chang HY, Yeh CY, Lee CT, Lin CC. A sleep apnea detection system based on a one-dimensional deep convolution neural network model using single-lead electrocardiogram. Sensors (Basel) 2020 Jul 26;20(15):4157 [FREE Full text] [doi: 10.3390/s20154157] [Medline: 32722630]
290. Sharan RV, Berkovsky S, Xiong H, Coiera E. ECG-derived heart rate variability interpolation and 1-D convolutional neural networks for detecting sleep apnea. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 20-24, 2020; Montreal, QC, Canada URL: https://doi.org/10.1109/embc44109.2020.9175998 [doi: 10.1109/embc44109.2020.9175998]
291. Urtnasan E, Park JU, Joo EY, Lee KJ. Identification of sleep apnea severity based on deep learning from a short-term normal ECG. J Korean Med Sci 2020 Dec 07;35(47):e399 [FREE Full text] [doi: 10.3346/jkms.2020.35.e399] [Medline: 33289367]
292. Jarchi D, Andreu-Perez J, Kiani M, Vysata O, Kuchynka J, Prochazka A, et al. Recognition of patient groups with sleep related disorders using bio-signal processing and deep learning. Sensors (Basel) 2020 May 02;20(9):2594 [FREE Full text] [doi: 10.3390/s20092594] [Medline: 32370185]
293. Li A, Chen S, Quan SF, Powers LS, Roveda JM. A deep learning-based algorithm for detection of cortical arousal during sleep. Sleep 2020 Dec 14;43(12):zsaa120 [FREE Full text] [doi: 10.1093/sleep/zsaa120] [Medline: 32556242]
294. Mashrur FR, Islam MS, Saha DK, Islam SR, Moni MA. SCNN: scalogram-based convolutional neural network to detect obstructive sleep apnea using single-lead electrocardiogram signals. Comput Biol Med 2021 Jul;134:104532 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104532] [Medline: 34102402]
295. Nasifoglu H, Erogul O. Obstructive sleep apnea prediction from electrocardiogram scalograms and spectrograms using convolutional neural networks. Physiol Meas 2021 Jun 29;42(6) [FREE Full text] [doi: 10.1088/1361-6579/ac0a9c] [Medline: 34116519]
296. Mukherjee D, Dhar K, Schwenker F, Sarkar R. Ensemble of deep learning models for sleep apnea detection: an experimental study. Sensors (Basel) 2021 Aug 11;21(16):5425 [FREE Full text] [doi: 10.3390/s21165425] [Medline: 34450866]
297. Urtnasan E, Joo EY, Lee KH. Ai-enabled algorithm for automatic classification of sleep disorders based on single-lead electrocardiogram. Diagnostics (Basel) 2021 Nov 05;11(11):2054 [FREE Full text] [doi: 10.3390/diagnostics11112054] [Medline: 34829400]
298. Yang Q, Zou L, Wei K, Liu G. Obstructive sleep apnea detection from single-lead electrocardiogram signals using one-dimensional squeeze-and-excitation residual group network. Comput Biol Med 2021 Dec 06;140:105124 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.105124] [Medline: 34896885]
299. Krasteva V, Ménétré S, Didon JP, Jekova I. Fully convolutional deep neural networks with optimized hyperparameters for detection of shockable and non-shockable rhythms. Sensors (Basel) 2020 May 19;20(10):2875 [FREE Full text] [doi: 10.3390/s20102875] [Medline: 32438582]

300. Isasi I, Irusta U, Aramendi E, Eftestøl T, Kramer-Johansen J, Wik L. Rhythm analysis during cardiopulmonary resuscitation using convolutional neural networks. Entropy (Basel) 2020 May 27;22(6):595 [FREE Full text] [doi: 10.3390/e22060595] [Medline: 33286367]
301. Miura K, Goto S, Katsumata Y, Ikura H, Shiraishi Y, Sato K, et al. Feasibility of the deep learning method for estimating the ventilatory threshold with electrocardiography data. NPJ Digit Med 2020;3:141 [FREE Full text] [doi: 10.1038/s41746-020-00348-6] [Medline: 33145437]
302. Kwon JM, Kim KH, Medina-Inojosa J, Jeon KH, Park J, Oh BH. Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. J Heart Lung Transplant 2020 Aug;39(8):805-814 [FREE Full text] [doi: 10.1016/j.healun.2020.04.009] [Medline: 32381339]
303. Wang L, Mu Y, Zhao J, Wang X, Che H. IGRNet: a deep learning model for non-invasive, real-time diagnosis of prediabetes through electrocardiograms. Sensors (Basel) 2020 Apr 30;20(9):2556 [FREE Full text] [doi: 10.3390/s20092556] [Medline: 32365875]
304. Ahmad Z, Khan NM. Multi-level stress assessment using multi-domain fusion of ECG signal. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Jul 20-24, 2020; Montreal, QC, Canada URL: https://doi.org/10.1109/embc44109.2020.9176590 [doi: 10.1109/embc44109.2020.9176590]
305. Hajeb - M S, Cascella A, Valentine M, Chon K. Deep neural network approach for continuous ECG - based automated external defibrillator shock advisory system during cardiopulmonary resuscitation. J Am Heart Assoc 2021 Mar 16;10(6):e019065 [FREE Full text] [doi: 10.1161/jaha.120.019065]
306. Jekova I, Krasteva V. Optimization of end-to-end convolutional neural networks for analysis of out-of-hospital cardiac arrest rhythms during cardiopulmonary resuscitation. Sensors (Basel) 2021 Jun 15;21(12):4150 [FREE Full text] [doi: 10.3390/s21124105] [Medline: 34203701]
307. Dunn AJ, ElRefai MH, Roberts PR, Coniglio S, Wiles BM, Zemkoho AB. Deep learning methods for screening patients' S-ICD implantation eligibility. Artif Intell Med 2021 Sep;119:102139 [FREE Full text] [doi: 10.1016/j.artmed.2021.102139] [Medline: 34531008]
308. Kwon JM, Jung MS, Kim KH, Jo YY, Shin JH, Cho YH, et al. Artificial intelligence for detecting electrolyte imbalance using electrocardiography. Ann Noninvasive Electrocardiol 2021 May;26(3):e12839 [FREE Full text] [doi: 10.1111/anec.12839] [Medline: 33719135]
309. Ozdemir MA, Ozdemir GD, Guren O. Classification of COVID-19 electrocardiograms by using hexaxial feature mapping and deep learning. BMC Med Inform Decis Mak 2021 May 25;21(1):170 [FREE Full text] [doi: 10.1186/s12911-021-01521-x] [Medline: 34034715]
310. Noor ST, Asad ST, Khan MM, Gaba GS, Al-Amri JF, Masud M. Predicting the risk of depression based on ECG using RNN. Comput Intell Neurosci 2021;2021:1299870 [FREE Full text] [doi: 10.1155/2021/1299870] [Medline: 34367269]
311. Chang DW, Lin CS, Tsao TP, Lee CC, Chen JT, Tsai CS, et al. Detecting digoxin toxicity by artificial intelligence-assisted electrocardiography. Int J Environ Res Public Health 2021 Apr 06;18(7):3839 [FREE Full text] [doi: 10.3390/ijerph18073839] [Medline: 33917563]
312. Lin C, Lee Y, Fang W, Lou Y, Kuo F, Lee C, et al. Deep learning algorithm for management of diabetes mellitus via electrocardiogram-based glycated hemoglobin (ECG-HbA1c): a retrospective cohort study. J Pers Med 2021 Jul 27;11(8):725 [FREE Full text] [doi: 10.3390/jpm11080725] [Medline: 34442369]
313. Baghersalimi S, Teijeiro T, Atienza D, Aminifar A. Personalized real-time federated learning for epileptic seizure detection. IEEE J Biomed Health Inform 2022 Feb;26(2):898-909 [FREE Full text] [doi: 10.1109/jbhi.2021.3096127]
314. Russell B, McDaid A, Toscano W, Hume P. Predicting fatigue in long duration mountain events with a single sensor and deep learning model. Sensors (Basel) 2021 Aug 12;21(16):5442 [FREE Full text] [doi: 10.3390/s21165442] [Medline: 34450884]
315. Bleijendaal H, Ramos LA, Lopes RR, Verstraelen TE, Baalman SW, Oudkerk Pool MD, et al. Computer versus cardiologist: is a machine learning algorithm able to outperform an expert in diagnosing a phospholamban p.Arg14del mutation on the electrocardiogram? Heart Rhythm 2021 Jan;18(1):79-87 [FREE Full text] [doi: 10.1016/j.hrthm.2020.08.021] [Medline: 32911053]
316. Lopes RR, Bleijendaal H, Ramos LA, Verstraelen TE, Amin AS, Wilde AA, et al. Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning: an application to phospholamban p.Arg14del mutation carriers. Comput Biol Med 2021 Apr;131:104262 [FREE Full text] [doi: 10.1016/j.compbiomed.2021.104262] [Medline: 33607378]
317. Lin C, Lin CS, Lee DJ, Lee CC, Chen SJ, Tsai SH, et al. Artificial intelligence-assisted electrocardiography for early diagnosis of thyrotoxic periodic paralysis. J Endocr Soc 2021 Sep 01;5(9):bvab120 [FREE Full text] [doi: 10.1210/jendso/bvab120] [Medline: 34308091]
318. Mazumder O, Banerjee R, Roy D, Mukherjee A, Ghose A, Khandelwal S, et al. Computational model for therapy optimization of wearable cardioverter defibrillator: shockable rhythm detection and optimal electrotherapy. Front Physiol 2021 Dec 10;12:787180 [FREE Full text] [doi: 10.3389/fphys.2021.787180] [Medline: 34955894]

319. He Z, Liu X, He H, Wang H. Dual attention convolutional neural network based on adaptive parametric ReLU for denoising ECG signals with strong noise. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630123 [doi: 10.1109/embc46164.2021.9630123]
320. Li WC, Yang CJ, Liu BT, Fang WC. A real-time affective computing platform integrated with AI system-on-chip design and multimodal signal processing system. In: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Nov 1-5, 2021; Mexico URL: https://doi.org/10.1109/embc46164.2021.9630979 [doi: 10.1109/embc46164.2021.9630979]
321. Kwon JM, Lee YR, Jung MS, Lee YJ, Jo YY, Kang DY, et al. Deep-learning model for screening sepsis using electrocardiography. Scand J Trauma Resusc Emerg Med 2021 Oct 03;29(1):145 [FREE Full text] [doi: 10.1186/s13049-021-00953-8] [Medline: 34602084]
322. Sarkar P, Lobmaier S, Fabre B, González D, Mueller A, Frasch MG, et al. Detection of maternal and fetal stress from the electrocardiogram with self-supervised representation learning. Sci Rep 2021 Dec 17;11(1):24146 [FREE Full text] [doi: 10.1038/s41598-021-03376-8] [Medline: 34921162]
323. Machine Learning Refined: Foundations, Algorithms, and Applications (2nd edition). Cambridge: Cambridge University Press; 2020. URL: https://www.cambridge.org/gr/academic/subjects/engineering/communications-and-signal-processing/machine-learning-refined-foundations-algorithms-and-applications-2nd-edition?format=HB
324. Jeon W, Ko G, Lee J, Lee H, Ha D, Ro HW. Deep learning with GPUs. In: Advances in Computers. Amsterdam: Elsevier Science; 2021. [doi: 10.1016/bs.adcom.2020.11.003]
325. Schutte AE, Kollias A, Stergiou GS. Blood pressure and its variability: classic and novel measurement techniques. Nat Rev Cardiol 2022 Apr 19:1-12 (forthcoming) [FREE Full text] [doi: 10.1038/s41569-022-00690-0] [Medline: 35440738]
326. Stergiou GS, Mukkamala R, Avolio A, Kyriakoulis KG, Mieke S, Murray A, European Society of Hypertension Working Group on Blood Pressure MonitoringCardiovascular Variability. Cuffless blood pressure measuring devices: review and statement by the European Society of Hypertension Working Group on Blood Pressure Monitoring and Cardiovascular Variability. J Hypertens 2022 Jun 17 [FREE Full text] [doi: 10.1097/HJH.0000000000003224] [Medline: 35708294]
327. Mukkamala R, Yavarimanesh M, Natarajan K, Hahn J, Kyriakoulis KG, Avolio AP, et al. Evaluation of the accuracy of cuffless blood pressure measurement devices: challenges and proposals. Hypertension 2021 Nov;78(5):1161-1167 [FREE Full text] [doi: 10.1161/hypertensionaha.121.17747]
328. Silverthorn DU, Michael J. Cold stress and the cold pressor test. Adv Physiol Educ 2013 Mar;37(1):93-96 [FREE Full text] [doi: 10.1152/advan.00002.2013] [Medline: 23471256]
329. Goldstein DS, Cheshire WP. Beat-to-beat blood pressure and heart rate responses to the Valsalva maneuver. Clin Auton Res 2017 Dec;27(6):361-367 [FREE Full text] [doi: 10.1007/s10286-017-0474-y] [Medline: 29052077]
330. Wanyan T, Honarvar H, Jaladanki SK, Zang C, Naik N, Somani S, et al. Contrastive learning improves critical event prediction in COVID-19 patients. Patterns (N Y) 2021 Dec 10;2(12):100389 [FREE Full text] [doi: 10.1016/j.patter.2021.100389] [Medline: 34723227]
331. Patel J, Shah S, Thakkar P, Kotecha K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Syst Applications 2015 Jan;42(1):259-268 [FREE Full text] [doi: 10.1016/j.eswa.2014.07.040]
332. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357 [FREE Full text] [doi: 10.1613/jair.953]
333. He H, Bai Y, Garcia EA, Shutao L. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008 Presented at: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); Jun 01-08, 2008; Hong Kong URL: https://doi.org/10.1109/ijcnn.2008.4633969 [doi: 10.1109/ijcnn.2008.4633969]
334. Sanabila HR, Kusuma I, Jatmiko W. Generative oversampling method (GenOMe) for imbalanced data on apnea detection using ECG data. In: Proceedings of the 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS). 2016 Presented at: 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS); Oct 15-16, 2016; Malang, Indonesia URL: https://doi.org/10.1109/icacsis.2016.7872805 [doi: 10.1109/icacsis.2016.7872805]
335. Rajesh KN, Dhuli R. Classification of imbalanced ECG beats using re-sampling techniques and AdaBoost ensemble classifier. Biomedical Signal Process Control 2018 Mar;41:242-254 [FREE Full text] [doi: 10.1016/j.bspc.2017.12.004]
336. Lin T, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020 Feb 1;42(2):318-327 [FREE Full text] [doi: 10.1109/tpami.2018.2858826]
337. Regression trees. In: Classification And Regression Trees. Milton Park, Abingdon-on-Thames, Oxfordshire, England, UK: Routledge; 1984.

338. Friedman J, Stuetzle W. Projection Pursuit Regression. J Am Statistical Assoc 1981 Dec;76(376):817-823 [FREE Full text] [doi: 10.1080/01621459.1981.10477729]
339. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. ArXiv 2014.
340. 340 TR. Regression shrinkage and selection via the lasso. J Royal Statistical Soc Series B (Methodological) 1996;58(1):267-288 [FREE Full text] [doi: 10.1111/j.2517-6161.1996.tb02080.x]
341. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010 May 15;26(10):1340-1347 [FREE Full text] [doi: 10.1093/bioinformatics/btq134] [Medline: 20385727]
342. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001 Oct 1;29(5):1189-1232 [FREE Full text] [doi: 10.1214/aos/1013203451]
343. Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations; Jun, 2016; San Diego, California URL: https://doi.org/10.18653/v1/n16-3020 [doi: 10.18653/v1/n16-3020]
344. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). 2017 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Oct 22-29, 2017; Venice, Italy URL: https://doi.org/10.1109/iccv.2017.74 [doi: 10.1109/iccv.2017.74]
345. GAL Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. 2017 Presented at: ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70; Aug 6 - 11, 2017; Sydney NSW Australia URL: https://proceedings.mlr.press/v70/gal17a.html
346. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929-1958.
347. Kupinski MA, Hoppin JW, Clarkson E, Barrett HH. Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques. J Opt Soc Am A Opt Image Sci Vis 2003 Mar;20(3):430-438 [FREE Full text] [doi: 10.1364/josaa.20.000430] [Medline: 12630829]
348. Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. 2015 Presented at: ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37; Jul 6 - 11, 2015; Lille France. [doi: 10.5555/3045118.3045290]
349. Ray TR, Choi J, Bandodkar AJ, Krishnan S, Gutruf P, Tian L, et al. Bio-integrated wearable systems: a comprehensive review. Chem Rev 2019 Apr 24;119(8):5461-5533 [FREE Full text] [doi: 10.1021/acs.chemrev.8b00573] [Medline: 30689360]
350. Kwak SS, Yoo S, Avila R, Chung HU, Jeong H, Liu C, et al. Skin-integrated devices with soft, holey architectures for wireless physiological monitoring, with applications in the neonatal intensive care unit. Adv Mater 2021 Nov;33(44):e2103974 [FREE Full text] [doi: 10.1002/adma.202103974] [Medline: 34510572]
351. Chung HU, Rwei AY, Hourlier-Fargette A, Xu S, Lee K, Dunne EC, et al. Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units. Nat Med 2020 Mar;26(3):418-429 [FREE Full text] [doi: 10.1038/s41591-020-0792-9] [Medline: 32161411]
352. Jeong H, Lee JY, Lee K, Kang YJ, Kim JT, Avila R, et al. Differential cardiopulmonary monitoring system for artifact-canceled physiological tracking of athletes, workers, and COVID-19 patients. Sci Adv 2021 May;7(20):eabg3092 [FREE Full text] [doi: 10.1126/sciadv.abg3092] [Medline: 33980495]
353. Lee K, Ni X, Lee JY, Arafa H, Pe DJ, Xu S, et al. Mechano-acoustic sensing of physiological processes and body motions via a soft wireless device placed at the suprasternal notch. Nat Biomed Eng 2020 Feb;4(2):148-158 [FREE Full text] [doi: 10.1038/s41551-019-0480-6] [Medline: 31768002]
354. Ni X, Ouyang W, Jeong H, Kim J, Tzaveils A, Mirzazadeh A, et al. Automated, multiparametric monitoring of respiratory biomarkers and vital signs in clinical and home settings for COVID-19 patients. Proc Natl Acad Sci U S A 2021 May 11;118(19):e2026610118 [FREE Full text] [doi: 10.1073/pnas.2026610118] [Medline: 33893178]
355. Chen SW, Wang SL, Qi XZ, Samuri SM, Yang C. Review of ECG detection and classification based on deep learning: coherent taxonomy, motivation, open challenges and recommendations. Biomedical Signal Process Control 2022 Apr;74:103493 [FREE Full text] [doi: 10.1016/j.bspc.2022.103493]
356. Hammad M, Kandala RN, Abdelatey A, Abdar M, Zomorodi - Moghadam M, Tan RS, et al. Automated detection of shockable ECG signals: a review. Inf Sci 2021 Sep;571:580-604 [FREE Full text] [doi: 10.1016/j.ins.2021.05.035]
357. Liu X, Wang H, Li Z, Qin L. Deep learning in ECG diagnosis: a review. Knowl Based Syst 2021 Sep;227:107187 [FREE Full text] [doi: 10.1016/j.knosys.2021.107187]

Abbreviations

AF: atrial fibrillation

AI: artificial intelligence
BIH: Beth Israel Hospital
BP: blood pressure
CNN: convolutional neural network
CPSC: China Physiological Signal Challenge
CVD: cardiovascular disease
DBP: diastolic blood pressure
DL: deep learning
DNN: deep neural network
ECG: electrocardiogram
GRU: gated recurrent unit
LSTM: long short-term memory
MIT: Massachusetts Institute of Technology
ML: machine learning
NN: neural network
PTB: Physikalisch Technische Bundesanstalt
ResNet: residual neural network
RNN: recurrent neural network
SBP: systolic blood pressure

Edited by C Lovis; submitted 03.04.22; peer-reviewed by H Turbe, S Fudickar; comments to author 08.05.22; revised version received 03.06.22; accepted 03.07.22; published 15.08.22.

Please cite as:
Petmezas G, Stefanopoulos L, Kilintzis V, Tzavelis A, Rogers JA, Katsaggelos AK, Maglaveras N
State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review
JMIR Med Inform 2022;10(8):e38454
URL: https://medinform.jmir.org/2022/8/e38454
doi:10.2196/38454
PMID:35969441

Original Paper

Predicting Abnormalities in Laboratory Values of Patients in the Intensive Care Unit Using Different Deep Learning Models: Comparative Study

Ahmad Ayad[1], MSc; Ahmed Hallawa[2], MSc; Arne Peine[2], Dr med, MHBA; Lukas Martin[2], Priv-Doz, Dr med, MHBA; Lejla Begic Fazlic[3], PhD; Guido Dartmann[3], Prof Dr-Ing; Gernot Marx[2], Prof Dr med, FRCA; Anke Schmeink[1], Prof Dr-Ing

[1]Chair of Information Theory and Data Analytics, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany

[2]Department of Intensive Care and Intermediate Care, University Hospital Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany

[3]Fachbereich Umweltplanung/Umwelttechnik - Fachrichtung Informatik, Trier University of Applied Sciences, Trier, Germany

Corresponding Author:
Ahmad Ayad, MSc
Chair of Information Theory and Data Analytics
Rheinisch-Westfälische Technische Hochschule Aachen
Kopernikusstraße 16
Aachen, 52074
Germany
Phone: 49 (241) 80 20750
Email: ahmad.ayad@inda.rwth-aachen.de

Abstract

Background: In recent years, the volume of medical knowledge and health data has increased rapidly. For example, the increased availability of electronic health records (EHRs) provides accurate, up-to-date, and complete information about patients at the point of care and enables medical staff to have quick access to patient records for more coordinated and efficient care. With this increase in knowledge, the complexity of accurate, evidence-based medicine tends to grow all the time. Health care workers must deal with an increasing amount of data and documentation. Meanwhile, relevant patient data are frequently overshadowed by a layer of less relevant data, causing medical staff to often miss important values or abnormal trends and their importance to the progression of the patient's case.

Objective: The goal of this work is to analyze the current laboratory results for patients in the intensive care unit (ICU) and classify which of these lab values could be abnormal the next time the test is done. Detecting near-future abnormalities can be useful to support clinicians in their decision-making process in the ICU by drawing their attention to the important values and focus on future lab testing, saving them both time and money. Additionally, it will give doctors more time to spend with patients, rather than skimming through a long list of lab values.

Methods: We used Structured Query Language to extract 25 lab values for mechanically ventilated patients in the ICU from the MIMIC-III and eICU data sets. Additionally, we applied time-windowed sampling and holding, and a support vector machine to fill in the missing values in the sparse time series, as well as the Tukey range to detect and delete anomalies. Then, we used the data to train 4 deep learning models for time series classification, as well as a gradient boosting–based algorithm and compared their performance on both data sets.

Results: The models tested in this work (deep neural networks and gradient boosting), combined with the preprocessing pipeline, achieved an accuracy of at least 80% on the multilabel classification task. Moreover, the model based on the multiple convolutional neural network outperformed the other algorithms on both data sets, with the accuracy exceeding 89%.

Conclusions: In this work, we show that using machine learning and deep neural networks to predict near-future abnormalities in lab values can achieve satisfactory results. Our system was trained, validated, and tested on 2 well-known data sets to ensure that our system bridged the reality gap as much as possible. Finally, the model can be used in combination with our preprocessing pipeline on real-life EHRs to improve patients' diagnosis and treatment.

(JMIR Med Inform 2022;10(8):e37658) doi:10.2196/37658

KEYWORDS

anomaly detection; DNN; time series classification; lab values; ICU; CNN; medical Informatics; EHR; machine learning; lightGBM

Introduction

Background

Machine learning and data analysis methods are used for diverse applications, such as anomaly detection [1], text classification [2], image segmentation [3], and time series forecasting [4]. One of the fields in which machine learning has become extremely popular recently is medicine. In medicine, there are now other application due to the improved availability of data. In particular, medical images [5] and electronic health records (EHRs) [6,7] represent prominent examples here. Much research has been done on medical images to detect diseases, such as pneumonia [8], which was driven by the advancements in computer vision. In addition, EHRs enabled the use of machine learning models to perform many tasks, such as predicting hospital length of stay [9] and mortality in septic patients [10]. In these studies, the authors used EHRs to train their machine learning models. However, EHRs have so much more data that with the right tools, they can support many valuable applications.

In this study, we consider the treatment of critically ill patients in the intensive care unit (ICU). Throughout the treatment of these patients, laboratory data are regularly gathered. Due to the substantial number of values to be monitored in the ICU, which sometimes can be more than 100 lab tests [11], important anomalies or trends may not be noticed. This can lead to suboptimal treatment strategies and complications in the patient's case. For example, early changes in lab values for patients with COVID-19 are important predictors of mortality [12]. The correct analysis of laboratory anomalies can direct treatment strategies, particularly in the early detection of potentially life-threatening cases. This should aid in resource allocation and save lives by allowing for timely intervention. Furthermore, health care workers spend 30%-50% of their time in front of computers and must deal with a mass of patient data [13,14]. Any savings in that time can free them to spend more time with patients.

Prior Work

Because of the recent availability of big data in the medical field, especially EHRs, there has been a growing interest in applying machine learning tools for medical applications. Working with medical data from EHRs can be quite challenging due to the inconsistent sampling of lab measurements, high frequency of missing values, and presence of noisy data. Additionally, there is no standardized way to process medical data before applying machine learning algorithms on them. Nevertheless, many authors have managed to process the data and apply machine learning algorithms for medical sequence modeling. Authors [15] have developed a masked, self-attention mechanism that uses positional encoding and dense interpolation strategies for incorporating temporal order. The authors trained and tested their model on the MIMIC-III data set and achieved better performance on them compared to recurrent neural networks (RNNs). The benchmarking tasks include predicting mortality (classification), length of stay (regression), phenotyping (multilabel classification), and decompensation (time series classification) [16]. Although the benchmarking tasks include a classification task, none of these tasks include lab values or the modeling of irregularly sampled sequences with large amounts of sparse data. The benchmark is created to compare different machine learning models on a specific type of medical data extracted from the MIMIC-III data set and covers only cover only 4 tasks. However, MIMIC-III has much more data that can allow for performing many more tasks like the one in this study.

There has also been some work that compares different approaches and machine learning algorithms for learning from irregularly sampled time series, which is mostly the case in medicine. For example, authors [17] compare modeling primitives that allow learning from the different forms of irregular time series, such as discretization, interpolation, recurrence, attention, and structural invariance. The authors discuss the pros and cons of each of these modeling primitives and the tasks for which they are suited. Another study [18] used a recurrence-based approach using specific versions of RNNs called gated recurrent units (GRUs) and discussed the advantages of using it instead of the other approaches. Additionally, authors [19] have proposed a system for early detection of sepsis using an interpolation-based method for data imputation followed by using temporal convolutional networks (TCNs) and dynamic time warping. The authors used a multitask gaussian process for multichannel data imputation and later used a TCN model to predict the probability of a sepsis diagnosis in the future. The authors proved that their proposed algorithm outperforms the state-of-the-art algorithm for sepsis detection. In contrast, we use a discretization-based approach followed by data imputation to convert the irregularly sampled time series to a regularly sampled one, as it provides an easy way to understand, debug, and implement a framework to deal with sensitive lab values that can be generalized effectively to other EHRs.

Goal of This Study

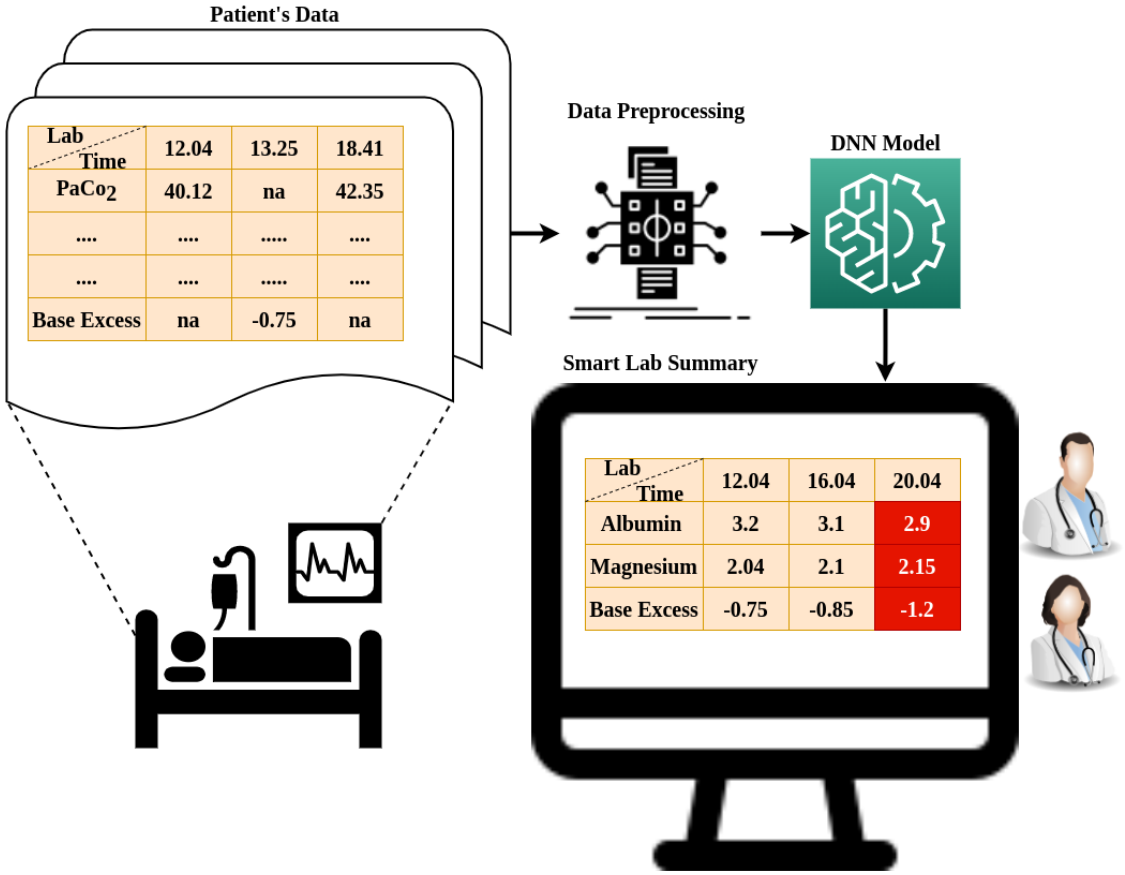
This work's objective is to analyze laboratory results (lab values) of patients in the ICU and classify which of these lab values are predicted to be out of the normal range soon (the next time these tests are done) and which are predicted to be normal. This allows health workers to focus on these laboratory values, their significance, their relation to the patient's current case, and their impact on the patient's future condition. This can potentially lead to reducing the length of the ICU stay and mortality [20]. Moreover, health care workers can focus future testing on these lab values and not waste time and resources on unnecessary tests that constitute approximately 50% of the tests ordered in the ICU [21]. Finally, it will allow the medical staff to reduce the time they need to check all the lab values and focus on the relevant ones, giving them more time to spend with patients [14].

Methods

Problem Definition

The task at hand is to predict which lab values will be normal and which will be abnormal in future, for a given period of ICU stay. The input data contain the patients' demographics and numerical lab values from the moment they were admitted till the end of their stay. The output is a binary vector, where each number represents the likelihood of a specific lab value to be abnormal (1) or normal (0) in the next 4 hours. Therefore, our problem is a "many to one" or a multilabel classification problem. Moreover, we have chosen the 4-hour time window because the majority of lab values found in MIMIC-III and eICU are recorded every 4 hours. Therefore, using this time step will introduce the least amount of data artifacts, especially considering that the changes in lab values are not noticeable for smaller time frames (like 1 hour). The same time window for lab values has been used by other authors [22]. Finally, the general diagram of the system is shown in Figure 1.

Figure 1. Overall abnormality detection system in practice. DNN: deep neural network.



Data and Cohort Definition

The data used to train, validate, and test the different prediction models are derived from the MIMIC-III database. It is a database that contains data from 31,532 unique ICU stays of patients who stayed within the ICUs at the Beth Israel Deaconess Medical Center [6] between 2001 and 2012. We also used data derived from the eICU Collaborative Research Database [7]. It is a multicenter database for critical care research created by The Philips eICU program. It contains data on 200,859 ICU stays from 335 ICUs units in the United States of America. In both databases, a unique ICU stay ID is associated with every unique ICU admission.

Our cohort focuses on mechanically ventilated patients in the ICU. This cohort is truly relevant these days because of the COVID-19 virus that caused a sharp increase in the number of patients in the ICU receiving mechanical ventilation. For these patients, it is vital to know which set of lab values have abnormal trends and focus on them, as it has a direct relation to how the case will develop [12]. The same cohort was used in a previous work focused on dynamically optimizing mechanical ventilation in critical care using reinforcement learning [22]. Using this cohort, we extracted 25,086 eICU and 11,943 MIMIC-III ICU stays with mechanical ventilation events. The duration of the ICU patients' stays ranges from 12 h to 72 h in 4-hour time steps. Patient demographics and clinical characteristics are shown in Table 1.

The input data consist of 3 demographic features (age, sex, weight) and 25 lab values (white blood cell count, $PaCO_2$, hemoglobin, etc). The lab values chosen are the most relevant to the mechanically ventilated patients, as shown by the medical team members from the university hospital of Rheinisch Westfälische Technische Hochschule (RWTH) Aachen in their previous work [22]. In Multimedia Appendix 1, the chosen features from the MIMIC-III and eICU data sets are listed along with their means and SDs.

The output is a binary vector of length 25. To convert numerical lab values to binary values, we used the reference ranges followed by the American College of Physicians [23]. Finally, the queries of Structured Query Language (SQL) used to extract the cohort data from both databases are included in the Git repository [24].

Table 1. Clinical and demographic properties of the study population [16].

Property	MIMIC-III data set	eICU data set
Number of ICUs[a]	5	335
Data acquisition timespan	2001-2012	2014-2015
Number of included patients (N)	11,443	23,699
Age (years), median (IQR)	66.9 (56.3-77.5)	65.0 (54-74)
Body weight in kg, mean (SD)	85.7 (18.1)	83.5 (22.0)
Sex, female, n(%)	4329 (36.3%)	10,546 (42%)
Sex, male, n (%)	7614 (63.7%)	14,540 (58%)
In-hospital mortality, %	11.1	13.2
LOS[b] in ICU (days), median (IQR)	3.1 (1.6-6.1)	3.0 (1.71-5.9)

[a]ICU: intensive care unit.

[b]LOS: length of stay.

Preprocessing

The patients' raw data extracted from the MIMIC-III and eICU data sets were very sparse and had several missing values. Therefore, it was necessary to perform preprocessing to prepare the data for the machine learning pipeline. First, the time-windowed sample-and-hold method was used to handle missing values. In this method, the data sample is held (repeated) until the next available data sample or the maximum hold time is reached. For each feature, we conducted a frequency analysis to determine how often a new measurement is produced. The counts of consecutive measurement time differences are obtained and when their cumulative sum exceeds a threshold, the first value where this occurs is taken as the hold time. When the feature's hold time exceeds this maximum, the data point is considered corrupted [25]. For the rest of the missing values, a k-nearest neighbor imputation with singular value decomposition and mean imputation were used [26]. Any ICU stay that had more than 50% missing data was discarded (occurrence <1% in the overall cohort) [22]. Finally, the Tukey range test was used to detect and delete outliers. The preprocessing steps are explained in detail in the Git repository [24].

Prediction System Overview

The overall system architecture used for predicting abnormalities in patients' lab values is shown in Figure 2. After performing the preprocessing steps explained earlier, the output time series will be separated into two main types: demographics and lab values. Each ICU stay will be split into multiple shorter sequences using the moving window technique. Figure 3 presents an example of an ICU stay of length L=11 (44 hours). Here, X_m represents the patient's input data vector at time step $m \in$  , and Y_m represents the patient's output binary vector. For a window size $W \in$  of 8, we have 3 subsequences extracted from the stay. For example, W_1 includes the input vectors [X_0:X_7] and the output binary vector Y_8. The process of the moving window is applied to ICU stays in the data sets (MIMIC-III, eICU). Then, the resulting subsequences are shuffled and used to train, validate, and test the different machine learning models that we have experimented with, as shown in Figure 2. This means the windowed subsequences from the same ICU stay can be distributed across the training, validation, and testing sets. Moreover, we experimented with different window sizes between $W=5$ and $W=10$ and chose the one that gave us the best results for all the models, as explained in the Results section.

We experimented with predicting the exact numerical lab values (regression problem) and then converting the predicted output to a binary vector after comparing the values with the normal ranges. The models were then trained to minimize the minimum squared error loss. The results were 10%-20% worse than those obtained when predicting the output binary vector directly and optimizing for the binary cross-entropy loss. Therefore, we selected this system model.

Figure 2. Overall system model used in our study when trained on the MIMIC-III data set and tested on the eICU data set. ICU: intensive care unit; Sigmoid is an activation function; L: lab value; t: time step.

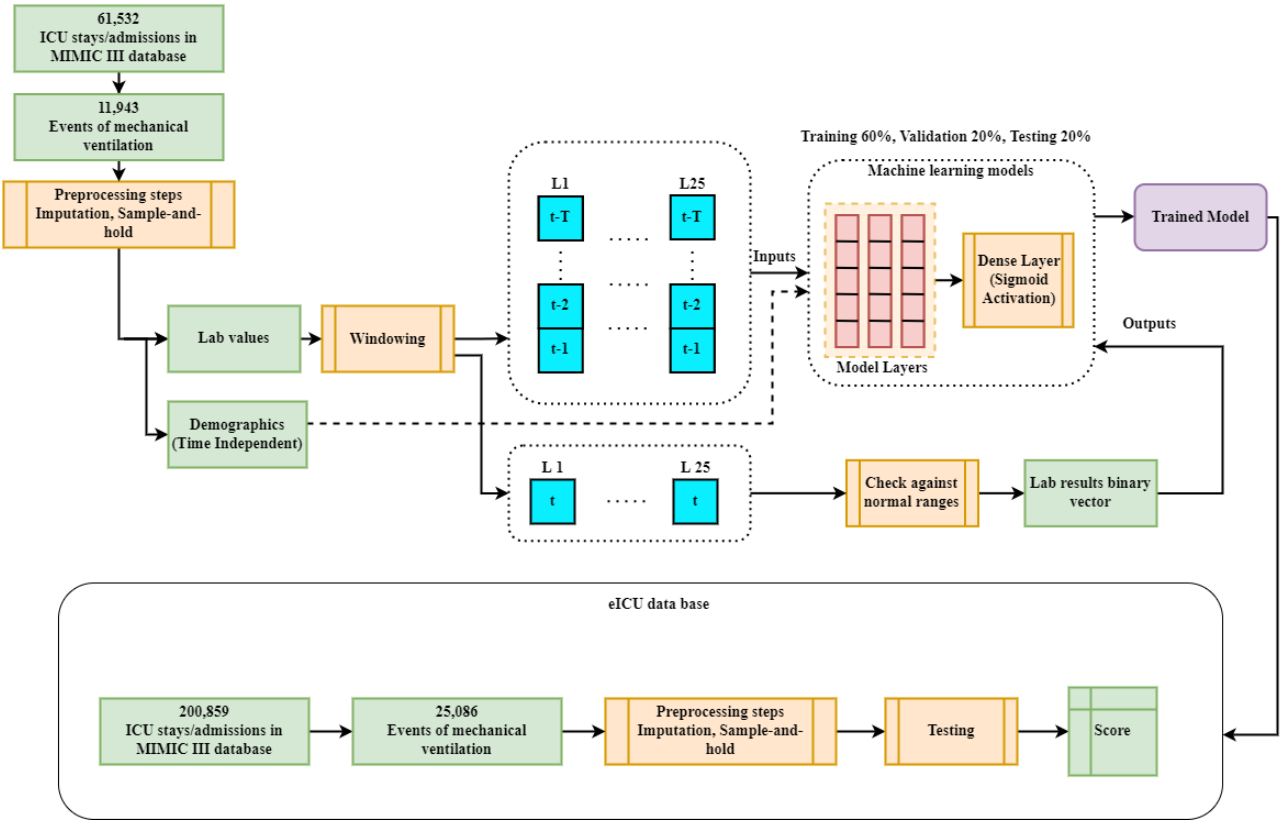
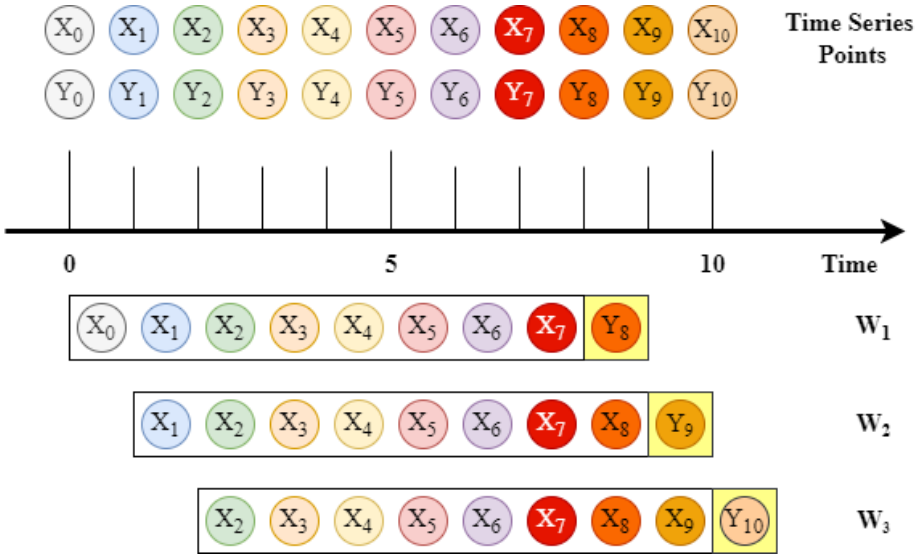


Figure 3. Moving window technique to extract sequences from intensive care unit stays. X and Y represent the input and output data respectively; W represents the windows extracted from the input sequences.



Prediction Models

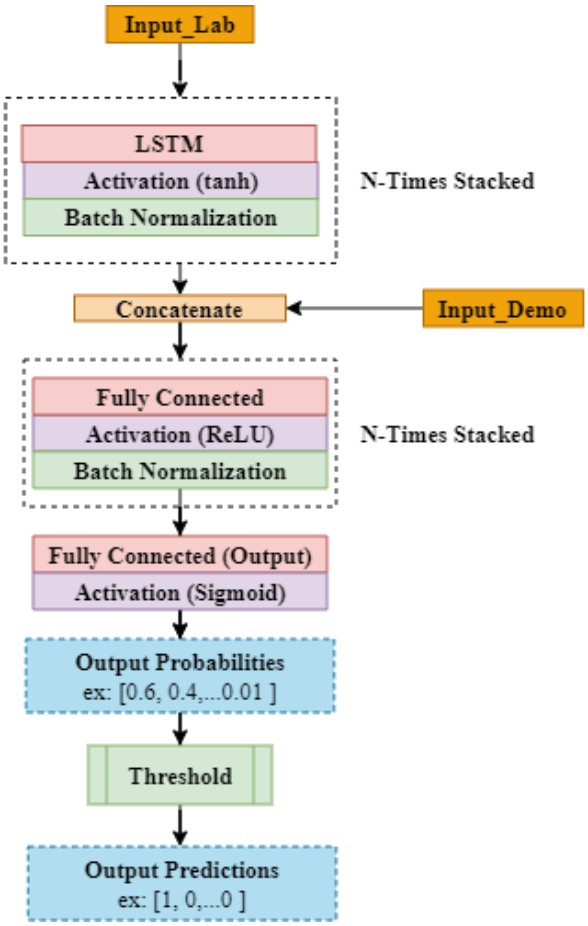
The goal of the prediction model in our scenario is to predict abnormalities in laboratory values for a given input sequence. The machine learning problem is a multilabel classification problem because multiple lab values are classified as normal or abnormal at the same time (multiclass) and more than 1 lab value can be abnormal at the same time (multilabel). We experimented with four current deep learning (DL) approaches: long short-term memory (LSTM), self-attention with time encoding (transformer architecture), convolutional neural network (CNN), and TCN. In the following subsections, each model architecture is discussed briefly. The models are explained in more detail in Multimedia Appendix 2 [2,27-39].

LSTM models

LSTM is a type of RNN that has the ability to learn from long sequences of data. A typical LSTM layer in a DL model consists of multiple LSTM cells. Another similar yet simpler cell structure is called GRU [4]. We experimented with both cell types in our model and chose LSTM because it performed better. The architecture used in our experiment is shown in Figure 4.

All the lab values will be input to the LSTM block to learn from the sequential data. Each LSTM block includes an LSTM layer, which has "tanh" as the built-in activation function. Then comes a batch normalization layer after the sequential data pass through the layers, and these data will be concatenated with the demographic features. The concatenated data will then go through a stack of fully connected layers ending with a last dense layer that has a sigmoid activation function. During forward propagation, the output probabilities will be compared to a threshold to produce the binary labels that are used to calculate the loss and other evaluation metrics.

Figure 4. LSTM architecture used in our experiments. LSTM: long short-term memory; ReLU: rectified linear unit; Tanh, ReLU and Sigmoid are activation functions.

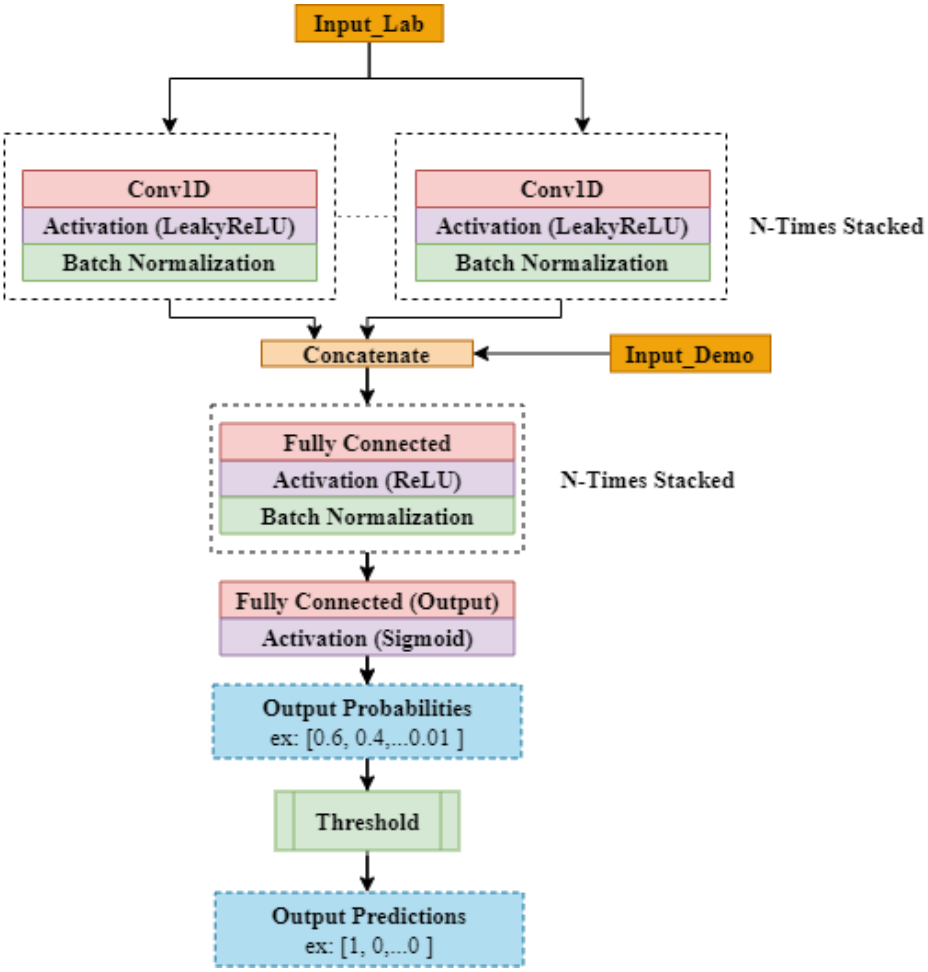


CNN models

CNNs learn to optimize their kernels to extract information from input data in a successive manner. Additionally, they work well on time series forecasting and classification problems [27], often outperforming LSTMs in terms of the total training time in a more computationally efficient manner [28]. In our case, we used a 1D multiple CNN (M-CNN), where the kernels (filters) move along the time axis performing convolution operations on all features. The kernel size defines how many time steps 1 kernel covers at any point in time.

Aside from the normal CNN that takes 1 input stream, we developed an architecture that takes 2 streams of the input sequences in parallel. Each stream will be processed with different filters. This ensures that we capture short-term dependencies in the sequences as well as long-term ones. The network architecture is shown in Figure 5.

Figure 5. Multiple convolutional neural network model architecture used in our experiments. Conv1D: 1D convolutional layer; LeakyReLU: leaky rectified linear unit; ReLU: rectified linear unit; Sigmoid, LeakyReLU, and ReLU are activation functions.

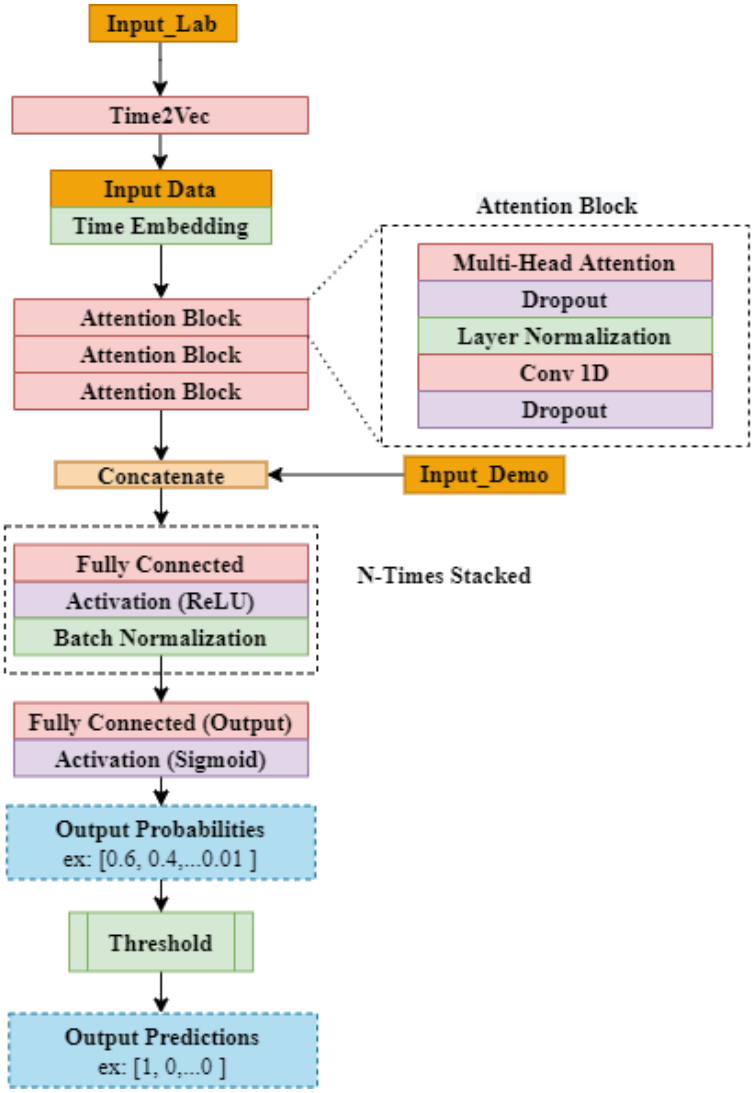


Transformer models

Transformers are a recent neural network architecture derived from the attention mechanism first proposed in an earlier study [29]. The mechanism was designed initially for translation tasks, which were earlier accomplished using RNNs.

Transformers typically use a collection of superimposed sinusoidal functions to represent the position of words in natural language processing tasks. However, in time series tasks, we need to attach the meaning of time to our input. Authors [30] have introduced a method where each input feature is represented as a linear component and a periodic component. The result at the end will be a learned vector representation of time steps that will be concatenated with the input data before the attention layers. The model architecture we developed is shown in Figure 6.

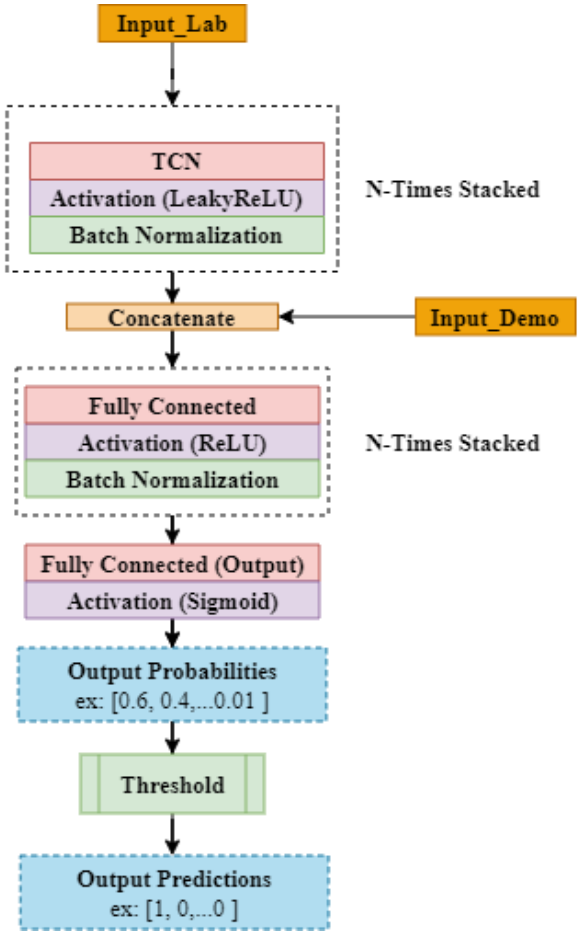
Figure 6. Transformer architecture used in our experiments. Conv1D: 1D convolutional layer; Time2Vec: time to vector transformation; ReLU: rectified linear unit; ReLU and Sigmoid are activation functions.



TCN models

TCNs were first introduced for video-based action segmentation [31]. Not long after that, they were used for sequence modeling tasks like the detection of sepsis [19]. A TCN differs from a conventional CNN in 2 ways; first, a TCN can take a sequence of any length and output a sequence of the same length using 0 padding; second, a TCN performs causal convolution. In general, TCNs are advantageous because they can be trained in parallel with less memory unlike RNNs. Additionally, they support variable length inputs and can easily replace any existing RNN. Figure 7 shows the TCN architecture that we designed and used in our experiments.

Figure 7. TCN architecture used in our experiments. LeakyReLU: leaky rectified linear unit; ReLU: rectified linear unit; TCN: temporal convolutional network; LeakyReLU, ReLU, and Sigmoid are activation functions.



Evaluation Metrics

In our work, we predicted the output binary vector of the future time step rather than the actual numerical lab values. We tried training the models as regression models predicting the actual numerical values and minimizing the minimum squared error. Then, we converted the predicted numerical output to binary vectors using the recommended ranges. However, we received better results when we treated the models as multilabel, multiclass classifiers predicting the binary vectors directly. Therefore, the evaluation metrics we used are binary accuracy, precision, recall, and F1 score.

Evaluation Setup

As we were predicting multiple lab values at the same time and all the classes were of equal importance, we used micro-averaging to calculate the accuracy, precision, recall, and F1 globally. These evaluation metrics were used to evaluate the models' training, validation, and testing. Additionally, to compare the models, the following points were followed: First, the models' architectures and hyperparameters were optimized using the Keras Tuner library [40] to ensure that the models performed at their best. Second, the models were trained to optimize the binary cross-entropy loss [41]. Third, early stopping was used to stop the model's training once the validation loss did not change by 0.01 for 10 consecutive epochs. This reduces the chances of model overfitting. Fourth, we set the seed for all the random processes during model training to ensure replicability of our results. Finally, we used the same threshold (TH=0.5) and same window size (sequence length=6) for all the models to ensure a fair comparison. We used 0 padding for sequences shorter than 6 time steps (ICU stay length<24 hours). Moreover, we implemented a gradient boosting–based method (LightGBM) for comparison with DL-based methods. LightGBM is one of the best performing non-DL–based algorithms that is shown to perform well on time series classification tasks [32].

We experimented with 2 approaches for training the models. In the first approach, we trained the models and validated them on the MIMIC-III data set. Then, we tested them on the MIMIC-III and eICU data sets, as shown in Figure 2. In the second approach, we trained and validated them on the eICU data set instead. Then, we tested them on the eICU and MIMIC-III data sets. Table 2 shows counts of the training, validation, and testing samples used in both methods from each data set (window size=6). The same cohort of patients was used in both cases, but eICU has much more patient data that led to a much bigger set than MIMIC-III. Finally, the model architectures and hyperparameters can be found on our Git repository [24] and in Multimedia Appendix 2.

Table 2. Sample counts for training, validation, and testing in both training methods.

Method	Number of training samples	Number of validation samples	Number of first testing samples	Number of second testing samples
#1	73,190 (MIMIC-III)	12,915 (MIMIC-III)	21,526 (MIMIC-III)	196,208 (eICU)
#2	166,776 (eICU)	29,431 (eICU)	49,052 (eICU)	86,106 (MIMIC-III)

Ethics Approval

Approval for data collection, processing, and release for the MIMIC-III database has been granted by the Institutional Review Boards of the Beth Israel Deaconess Medical Center (Boston, United States) and Massachusetts Institute of Technology (Cambridge, United States). Approval for data collection, processing, and release for the eICU database has been granted by the eICU research committee and exempt from Institutional Review Board approval. All data were processed using the computational infrastructure at the RWTH Aachen University and the University Hospital at RWTH Aachen in accordance with European Union data protection laws.

Results

In Figures 8, 9 and 10, we report the validation loss, F1 score, and accuracy of the different models during training, respectively. The models' names ending with "mimic" indicate that they were trained on the MIMIC-III data set and those ending in "eicu" refer to the models trained on the eICU data set. Moreover, because of the early stopping used during training, some models stopped training before others. Thus, their metrics are constant after the stopping point.

In Tables 3 and 4, we report the testing accuracy, recall, precision, and F1 scores of the different models. All the results were averaged over all the lab values and the testing samples.

As we expect our system to run continuously on huge amounts of data in hospitals, we want the performance of the chosen model to be good enough to meet such demands. Therefore, we measured the models' inference times. Experiments were run on a computer with an Intel(R) Core i9-9900K processor (Intel Corporation) running at 3.60 GHz using a 32-GB DDR4 RAM and Nvidia GTX 1080ti graphics processing unit (Nvidia Corporation), running Ubuntu (version 20.04, Canonical Ltd), Python (version 3.8, Python Software Foundation), and TensorFlow (version 2.6, Google Brain). Table 5 reports the inference time for each model on a whole batch (batch size=128 samples).

Figure 8. Validation loss of the different models. LSTM: long short-term network; M-CNN: multiple convolutional neural network; TCN: temporal convolutional network; Val.: validation; ICU: intensive care unit.

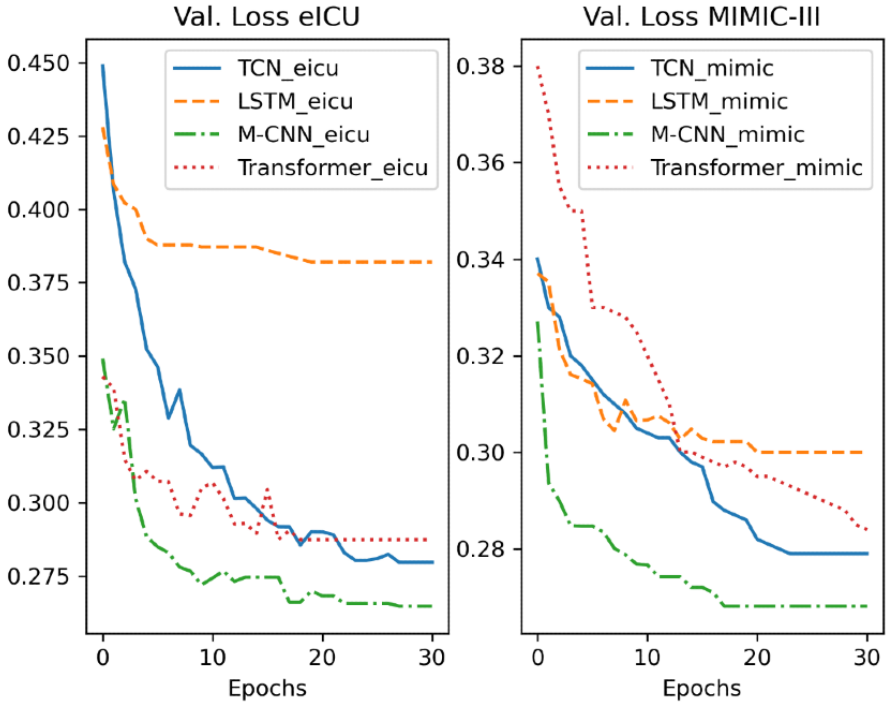


Figure 9. Validation F1 score of the different models. LSTM: long short-term network; M-CNN: multiple convolutional neural network; TCN: temporal convolutional network; Val.: validation; ICU: intensive care unit.

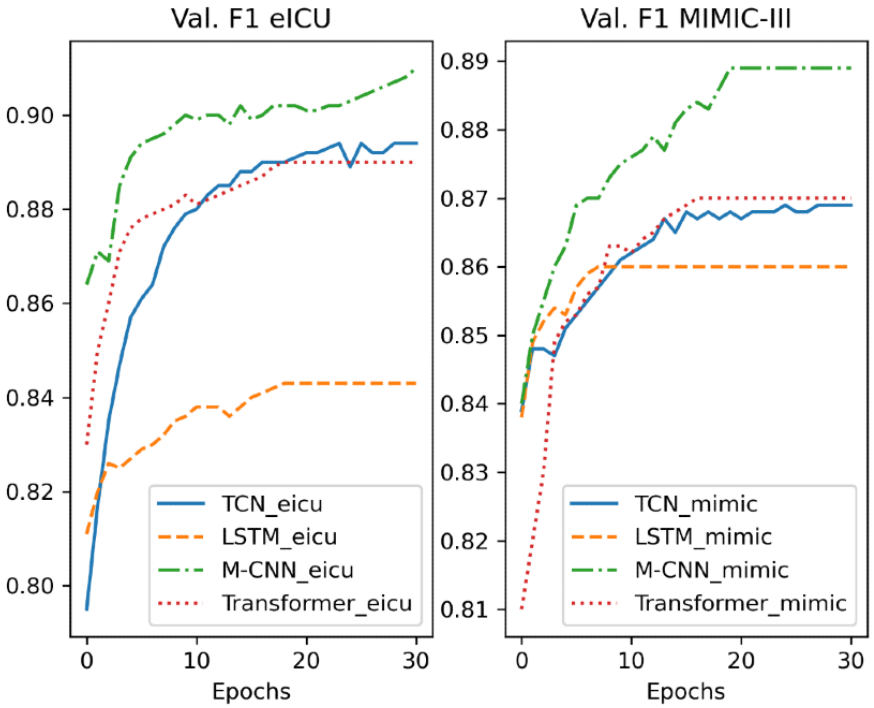


Figure 10. Validation accuracy of the different models. LSTM: long short-term network; M-CNN: multiple convolutional neural network; TCN: temporal convolutional network; Val.: validation; ICU: intensive care unit.

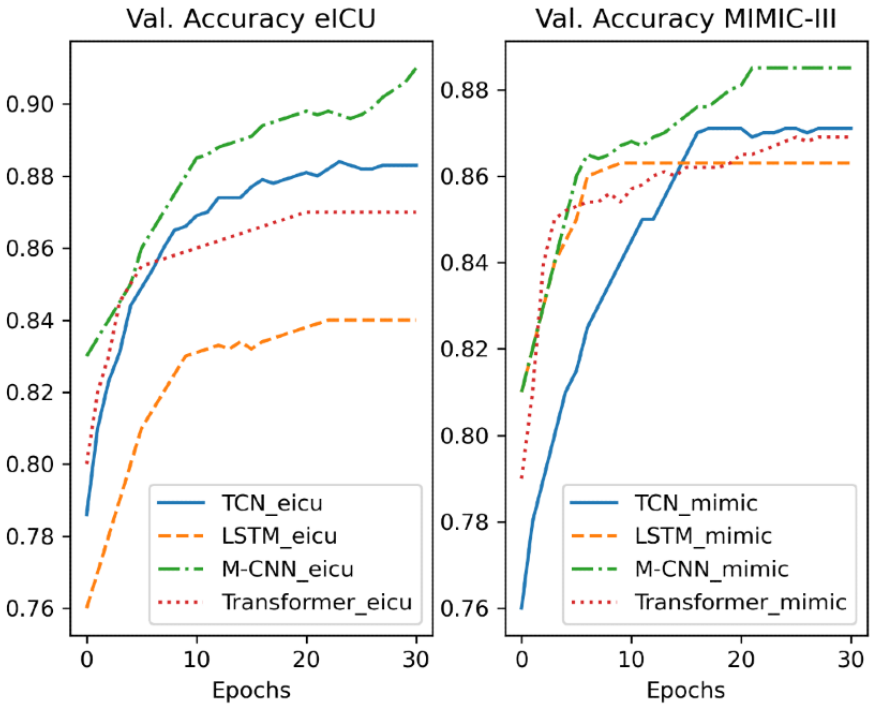


Table 3. Testing results for the different models over all lab values (micro-average) on the MIMIC-III data set[a].

Training data set and model	Accuracy	Precision	Recall	F1 score
MIMIC-III				
LSTM[b]	0.85	0.83	0.87	0.85
CNN[c]	0.86	0.84	0.85	0.84
M-CNN[d]	0.88	0.87	0.89	0.88
Transformer	0.86	0.88	0.81	0.84
TCN[e]	0.86	0.87	0.85	0.86
LightGBM[f]	0.83	0.82	0.76	0.78
eICU				
LSTM	0.8	0.79	0.81	0.8
CNN	0.85	0.86	0.83	0.84
M-CNN	0.87	0.88	0.86	0.87
Transformer	0.86	0.86	0.84	0.85
TCN	0.83	0.82	0.84	0.83
LightGBM	0.82	0.77	0.78	0.77

[a]The models listed under MIMIC-III were trained on the MIMIC-III data set and those under eICU were trained on the eICU data set.

[b]LSTM: long short-term memory.

[c]CNN: convolutional neural network.

[d]M-CNN: multiple convolutional neural network.

[e]TCN: temporal convolutional network.

[f]LightGBM: gradient boosting–based method.

Table 4. Testing results for the different models over all lab values (micro-average) on the eICU data set[a].

Training data set and model	Accuracy	Precision	Recall	F1 score
MIMIC-III				
LSTM[b]	0.79	0.81	0.8	0.8
CNN[c]	0.78	0.8	0.8	0.8
M-CNN[d]	0.8	0.8	0.83	0.81
Transformer	0.75	0.82	0.69	0.75
TCN[e]	0.71	0.74	0.72	0.73
LightGBM[f]	0.75	0.78	0.75	0.76
eICU				
LSTM	0.82	0.85	0.83	0.84
CNN	0.85	0.86	0.83	0.84
M-CNN	0.89	0.9	0.91	0.9
Transformer	0.86	0.87	0.88	0.87
TCN	0.89	0.88	0.89	0.89
LightGBM	0.82	0.77	0.78	0.77

[a]The models under MIMIC-III were trained on the MIMIC-III data set and those under eICU were trained on the eICU data set.
[b]LSTM: long short-term memory.
[c]CNN: convolutional neural network.
[d]M-CNN: multiple convolutional neural network.
[e]TCN: temporal convolutional network.
[f]LightGBM: gradient boosting–based method.

Table 5. Inference time for the different models.

Model name	Average inference time/batch
LSTM[a]	654 ms
CNN[b]	220 ms
M-CNN[c]	285 ms
TCN[d]	854 ms
Transformer	598 ms
LightGBM[e]	121 ms

[a]LSTM: long short-term memory.
[b]CNN: convolutional neural network.
[c]M-CNN: multiple convolutional neural network.
[d]TCN: temporal convolutional network.
[e]LightGBM: gradient boosting–based method.

Discussion

In this work, we developed an end-to-end system to extract and process lab results from EHRs and applied various machine learning algorithms to determine which lab values will be out of range in the next 4 hours with satisfactory results. This enables medical staff to focus on these lab values that can lead to improvements in overall patient diagnosis and treatment. Additionally, it can help reduce the time and cost wasted on irrelevant lab tests. The following steps were taken to reach this goal: First, we used SQL queries to extract the relevant patient data following our cohort from MIMIC-III and eICU data sets. Second, we used the time-windowed sample-and-hold method alongside k-nearest neighbor imputation with mean imputation and singular value decomposition to fill missing values. Moreover, we used the Tukey range test to detect anomalies and delete them. Third, we experimented with non-DL methods like LightGBM as well as 4 DL algorithms for time series classification. The DL-based method stacks models through mapping and processing functions between the models, using

gradient descent or momentum methods to optimize fit. Gradient boosting methods like LightGBM iteratively fit models to error terms and average results within a generalized linear modeling framework using base learner models at each iteration, introducing a penalty term into the base learner models. Finally, we trained and tested our algorithms on 2 of the well-known EHR data sets, MIMIC-III and eICU. Cross-validating our algorithms on these 2 data sets ensures not only a broader performance comparison, but also helps analyze how far the different algorithms can generalize on new unseen data.

A deeper analysis of the training results of the different DL-based models (Figures 8, 9 and 10) revealed that the M-CNN model trained on the eICU data set yielded better results at the end of the training than any other model. Additionally, we can see that the performance of both the TCN and transformer model improved significantly when trained on more data (eICU data set). This can be better understood from the results in Tables 3 and 4. First, the models trained on the eICU data set generalized better on data that they had not seen before from both the data sets. This is because the models had more data to train on, so they could see more variations and cases that they learned. On the other hand, the models trained on the MIMIC-III data set (43% the size of eICU training samples) performed well on the testing samples from MIMIC-III but performed much worse on the testing samples from eICU. Second, the M-CNN model performed the best in terms of almost all the evaluation metrics in both training methods. CNN models perform well on many sequenced modeling tasks, often outperforming RNN architectures like LSTM or GRU. Additionally, CNN-based models have the least number of trainable parameters out of the different DL-based methods and occupy the least memory, making them perform better on data sets with small amounts of training data. On the other hand, standard CNNs can only work with fixed-size inputs and usually focus on data elements that are in immediate proximity due to their static convolutional filter size. However, combining multiple CNN models helps increase the accuracy further by applying convolutions with multiple filter sizes and combining the outputs to give a more robust prediction. Moreover, in our case, we chose a static, relatively short input sequence length, thus mitigating the issue of long, variable length sequences. In case of long, variable length input sequences, a TCN will be a better candidate. A TCN employs techniques like multiple layers of dilated convolutions and padding of input sequences to handle different sequence lengths and detect dependencies between items that are not next to each other but are positioned on different places in a sequence. Furthermore, more complicated architectures like transformers and TCNs with many more trainable parameters would perform better if they had access to more data, which is often an issue in the medical field because of the scarcity of available training data. Therefore, M-CNN architectures are desirable for modeling medical time series data with static lengths and relatively short lengths like lab values requiring relatively smaller training data sets. Moreover, the M-CNN architecture can generalize well on unseen data when trained well, considering integrated measures for reducing overfitting during model training. An interesting fact is that despite not outperforming the M-CNN model, lightGBM performed as well (sometimes better) as some other DL-based approaches while requiring much less training time. Non-DL–based approaches can model problems with much less training data but require hand-crafted features and are very sensitive to outliers and variation in data. Further, removing seasonality is often needed when dealing with time series data. Finally, we can see that the LightGBM model is the fastest in terms of the inference time according to Table 5, followed by the CNN model, which is the fastest among the DL-based models. The M-CNN model, despite outperforming the regular CNN model, is 29% slower in terms of the inference time, which is expected as the model has more parameters.

Overall, our comprehensive analysis shows the advantage of using DL models for classifying future abnormalities in lab values for patients in the ICU. Although we tested our algorithms on 2 of the most used EHR data sets, further testing is needed to assess the performance of the full pipeline on other EHRs, including the preprocessing steps and how well the tuned hyperparameters of the machine learning models will generalize. Nevertheless, we believe this study can help other researchers trying to use machine learning in modeling medical time series problems.

Acknowledgments

This work is funded by the European Institute of Innovation & Technology (grant EIT-Health 19549). The funding institution of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the paper.

Authors' Contributions

AA, AH, AP, and LM conceived the idea. AA and AH performed data extraction. AS, GD, and GM provided methodological inputs. LF worked on the data cohort and SQL queries. AS and AH reviewed the mathematical analysis. AA and AH worked on the figures, tables, and manuscript writing. LM had full access to all data in the study. All authors read and approved the final submitted manuscript.

Conflicts of Interest

AP, GM, LM, AS, and GD are cofounders of Clinomic GmbH. AP and LM are chief executive officers of Clinomic GmbH. GM is the senior medical advisor, and GD and AS are scientific advisors in Clinomic GmbH. All remaining authors declare that they have no conflict of interests.

Multimedia Appendix 1
Statistical properties of the input features from MIMIC-III and eICU data sets.
[PDF File (Adobe PDF File), 84 KB - medinform_v10i8e37658_app1.pdf]

Multimedia Appendix 2
Details of the models used.
[PDF File (Adobe PDF File), 648 KB - medinform_v10i8e37658_app2.pdf]

References

1. Ayad A, Zamani A, Schmeink A, Dartmann G. Design and implementation of a hybrid anomaly detection system for IoT. In: 2019 Sixth International Conference on Internet of Things: Systems, Management and Security (IOTSMS).: IEEE; 2019 Oct Presented at: Sixth International Conference on Internet of Things: Systems, Management and Security; October 22-25, 2019; Granada, Spain p. 1-6. [doi: 10.1109/IOTSMS48152.2019.8939206]
2. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
3. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021 Feb;18(2):203-211. [doi: 10.1038/s41592-020-01008-z] [Medline: 33288961]
4. Lim B, Zohren S. Time-series forecasting with deep learning: a survey. Philos Trans A Math Phys Eng Sci 2021 Apr;379(2194):20200209. [doi: 10.1098/rsta.2020.0209] [Medline: 33583273]
5. Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, et al. VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. ArXiv Preprint posted online on Dec 30, 2020 [FREE Full text] [doi: 10.48550/arXiv.2012.15029]
6. 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;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
7. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep;5:180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
8. Kong L, Cheng J. Based on improved deep convolutional neural network model pneumonia image classification. PLoS One 2021 Nov;16(11):e0258804 [FREE Full text] [doi: 10.1371/journal.pone.0258804] [Medline: 34735483]
9. Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. Int J Cardiol 2019 Aug;288:140-147. [doi: 10.1016/j.ijcard.2019.01.046] [Medline: 30685103]
10. Perng J, Kao I, Kung C, Hung S, Lai Y, Su C. Mortality prediction of septic patients in the emergency department based on machine learning. J Clin Med 2019 Nov;8(11):1906 [FREE Full text] [doi: 10.3390/jcm8111906] [Medline: 31703390]
11. Frassica JJ. Frequency of laboratory test utilization in the intensive care unit and its implications for large-scale data collection efforts. J Am Med Inform Assoc 2005 Mar;12(2):229-233 [FREE Full text] [doi: 10.1197/jamia.M1604] [Medline: 15561793]
12. Kiss S, Gede N, Hegyi P, Németh D, Földi M, Dembrovszky F, et al. Early changes in laboratory parameters are predictors of mortality and ICU admission in patients with COVID-19: a systematic review and meta-analysis. Med Microbiol Immunol 2021 Feb;210(1):33-47 [FREE Full text] [doi: 10.1007/s00430-020-00696-w] [Medline: 33219397]
13. Butler R, Monsalve M, Thomas GW, Herman T, Segre AM, Polgreen PM, et al. Estimating time physicians and other health care workers spend with patients in an intensive care unit using a sensor network. Am J Med 2018 Aug;131(8):972.e9-972.e15. [doi: 10.1016/j.amjmed.2018.03.015] [Medline: 29649458]
14. Clinical artificial intelligence improving healthcare. eit Health. URL: https://eithealth.eu/product-service/clinical-artificial-intelligence-improving-healthcare [accessed 2022-04-20]
15. Song H, Rajan D, Thiagarajan J, Spanias A. Attend and diagnose: clinical time series analysis using attention models. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California, United States: AAAI Press; 2018 Apr Presented at: Thirty-Second AAAI Conference on Artificial Intelligence; February 2-7, 2018; New Orleans, LA, United States. [doi: 10.1609/aaai.v32i1.11635]
16. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019 Jun;6(1):96 [FREE Full text] [doi: 10.1038/s41597-019-0103-9] [Medline: 31209213]
17. Shukla SN, Marlin BM. A survey on principles, models and methods for learning from irregularly sampled time series. ArXiv Preprint posted online on Nov 30, 2020 [FREE Full text] [doi: 10.48550/arXiv.2012.00168]
18. Weerakody PB, Wong KW, Wang G, Ela W. A review of irregular time series data handling with gated recurrent neural networks. Neurocomputing 2021 Jun;441:161-178. [doi: 10.1016/j.neucom.2021.02.046]

19. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping. Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR 2019 Aug;106:2-26 [FREE Full text]
20. Tyler PD, Du H, Feng M, Bai R, Xu Z, Horowitz GL, et al. Assessment of intensive care unit laboratory values that differ from reference ranges and association with patient mortality and length of stay. JAMA Netw Open 2018 Nov;1(7):e184521 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.4521] [Medline: 30646358]
21. Mikhaeil M, Day AG, Ilan R. Non-essential blood tests in the intensive care unit: a prospective observational study. Can J Anaesth 2017 Mar;64(3):290-295. [doi: 10.1007/s12630-016-0793-9] [Medline: 28000153]
22. Peine A, Hallawa A, Bickenbach J, Dartmann G, Fazlic LB, Schmeink A, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. NPJ Digit Med 2021 Feb;4(1):32 [FREE Full text] [doi: 10.1038/s41746-021-00388-6] [Medline: 33608661]
23. ACP Internal Medicine Meeting. Reference ranges. URL: https://annualmeeting.acponline.org/educational-program/handouts/reference-ranges-table [accessed 2021-03-05]
24. Ayad A, Hallawa A, Schmeink A. Lab values abnormality detection (AI-LAD). a-ayad / AI_LAD. 2022. URL: https://github.com/a-ayad/AI_LAD [accessed 2022-01-25]
25. Mitra S. Digital Signal Processing: A Computer Based Approach. Europe: McGraw-Hill Education; 2010.
26. Salgado CM, Azevedo C, Proença H, Vieira SM. Missing data. In: Secondary Analysis of Electronic Health Records. Cham, Switzerland: Springer; Sep 2016.
27. Shi X, Huang G, Hao X, Yang Y, Li Z. A synchronous prediction model based on multi-channel CNN with moving window for coal and electricity consumption in cement calcination process. Sensors (Basel) 2021 Jun;21(13):4284 [FREE Full text] [doi: 10.3390/s21134284] [Medline: 34201548]
28. Bangyal WH, Qasim R, Rehman NU, Ahmad Z, Dar H, Rukhsar L, et al. Detection of fake news text classification on COVID-19 using deep learning approaches. Comput Math Methods Med 2021 Nov;2021:5514220 [FREE Full text] [doi: 10.1155/2021/5514220] [Medline: 34819990]
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems.: Curran Associates, Inc; 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA, United States URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
30. Kazemi M, Goel R, Eghbali S, Ramanan J, Sahota J, Thakur S, et al. Time2vec: learning a vector representation of time. ArXiv Preprint posted online on Jul 11, 2019 [FREE Full text]
31. Lea C, Vidal R, Reiter A, Hager GD. Temporal convolutional networks: a unified approach to action segmentation. In: Hua G, Jégou H, editors. Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science. Cham, Switzerland: Springer; Nov 2016.
32. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the Advances in Neural Information Processing Systems.: Curran Associates, Inc; 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA, United States URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
33. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
34. Li X, Chen S, Hu X, Yang J. Understanding the disharmony between dropout and batch normalization by variance shift. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).: IEEE; 2019 Jun Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 15-20, 2019; Long Beach, CA, United States. [doi: 10.1109/CVPR.2019.00279]
35. Matsugu M, Mori K, Mitari Y, Kaneda Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw 2003 Jun;16(5-6):555-559. [doi: 10.1016/S0893-6080(03)00115-1] [Medline: 12850007]
36. Jiang Y, Chen L, Zhang H, Xiao X. Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. PLoS One 2019 Mar;14(3):e0214587 [FREE Full text] [doi: 10.1371/journal.pone.0214587] [Medline: 30925170]
37. Wang H, Zhao J, Zhao H, Li H, Wang J. CL-ACP: a parallel combination of CNN and LSTM anticancer peptide recognition model. BMC Bioinformatics 2021 Oct;22(1):512 [FREE Full text] [doi: 10.1186/s12859-021-04433-9] [Medline: 34670488]
38. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the 30th International Conference on Machine Learning. 2013 Jun Presented at: 30th International Conference on Machine Learning; June 17-19, 2013; Atlanta, Georgia, United States.
39. Yan J, Mu L, Wang L, Ranjan R, Zomaya AY. Temporal convolutional networks for the advance prediction of ENSO. Sci Rep 2020 May;10(1):8055 [FREE Full text] [doi: 10.1038/s41598-020-65070-5] [Medline: 32415130]
40. Keras tuner library. keras-team/keras-tuner. 2022. URL: https://github.com/keras-team/keras-tuner [accessed 2021-10-09]
41. Murphy KP. Machine Learning: A Probabilistic Perspective. United States: MIT Press; 2013.

Abbreviations

CNN: convolutional neural network
DL: deep learning
EHR: electronic health record
GRU: gated recurrent unit
ICU: intensive care unit
LSTM: long short-term memory
M-CNN: multiple convolutional neural network
RNN: recurrent neural network
RWTH: Rheinisch Westfälische Technische Hochschule
SQL: Structured Query Language
TCN: temporal convolutional network

Edited by C Lovis; submitted 01.03.22; peer-reviewed by H Turbe, E Sükei; comments to author 11.04.22; revised version received 05.06.22; accepted 12.06.22; published 24.08.22.

Please cite as:
Ayad A, Hallawa A, Peine A, Martin L, Fazlic LB, Dartmann G, Marx G, Schmeink A
Predicting Abnormalities in Laboratory Values of Patients in the Intensive Care Unit Using Different Deep Learning Models: Comparative Study
JMIR Med Inform 2022;10(8):e37658
URL: https://medinform.jmir.org/2022/8/e37658
doi:10.2196/37658
PMID:36001363

Original Paper

Predicting Readmission Charges Billed by Hospitals: Machine Learning Approach

Deepika Gopukumar[1], PhD; Abhijeet Ghoshal[2], PhD; Huimin Zhao[3], PhD

[1]Department of Health and Clinical Outcomes Research, School of Medicine, Saint Louis University, St.Louis, MO, United States

[2]Department of Business Administration, Gies College of Business, University of Illinois Urbana-Champaign, Champaign, IL, United States

[3]Sheldon B Lubar College of Business, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

Corresponding Author:
Deepika Gopukumar, PhD
Department of Health and Clinical Outcomes Research
School of Medicine
Saint Louis University
SALUS Center, 3545 Lafayette Ave., 4th floor, Room 409 B
St.Louis, MO, 63110
United States
Phone: 1 3149779300
Email: deepika.gopukumar@health.slu.edu

Abstract

Background: The Centers for Medicare and Medicaid Services projects that health care costs will continue to grow over the next few years. Rising readmission costs contribute significantly to increasing health care costs. Multiple areas of health care, including readmissions, have benefited from the application of various machine learning algorithms in several ways.

Objective: We aimed to identify suitable models for predicting readmission charges billed by hospitals. Our literature review revealed that this application of machine learning is underexplored. We used various predictive methods, ranging from glass-box models (such as regularization techniques) to black-box models (such as deep learning–based models).

Methods: We defined readmissions as readmission with the same major diagnostic category (RSDC) and all-cause readmission category (RADC). For these readmission categories, 576,701 and 1,091,580 individuals, respectively, were identified from the Nationwide Readmission Database of the Healthcare Cost and Utilization Project by the Agency for Healthcare Research and Quality for 2013. Linear regression, lasso regression, elastic net, ridge regression, eXtreme gradient boosting (XGBoost), and a deep learning model based on multilayer perceptron (MLP) were the 6 machine learning algorithms we tested for RSDC and RADC through 10-fold cross-validation.

Results: Our preliminary analysis using a data-driven approach revealed that within RADC, the subsequent readmission charge billed per patient was higher than the previous charge for 541,090 individuals, and this number was 319,233 for RSDC. The top 3 major diagnostic categories (MDCs) for such instances were the same for RADC and RSDC. The average readmission charge billed was higher than the previous charge for 21 of the MDCs in the case of RSDC, whereas it was only for 13 of the MDCs in RADC. We recommend XGBoost and the deep learning model based on MLP for predicting readmission charges. The following performance metrics were obtained for XGBoost: (1) RADC (mean absolute percentage error [MAPE]=3.121%; root mean squared error [RMSE]=0.414; mean absolute error [MAE]=0.317; root relative squared error [RRSE]=0.410; relative absolute error [RAE]=0.399; normalized RMSE [NRMSE]=0.040; mean absolute deviation [MAD]=0.031) and (2) RSDC (MAPE=3.171%; RMSE=0.421; MAE=0.321; RRSE=0.407; RAE=0.393; NRMSE=0.041; MAD=0.031). The performance obtained for MLP-based deep neural networks are as follows: (1) RADC (MAPE=3.103%; RMSE=0.413; MAE=0.316; RRSE=0.410; RAE=0.397; NRMSE=0.040; MAD=0.031) and (2) RSDC (MAPE=3.202%; RMSE=0.427; MAE=0.326; RRSE=0.413; RAE=0.399; NRMSE=0.041; MAD=0.032). Repeated measures ANOVA revealed that the mean RMSE differed significantly across models with $P<.001$. Post hoc tests using the Bonferroni correction method indicated that the mean RMSE of the deep learning/XGBoost models was statistically significantly ($P<.001$) lower than that of all other models, namely linear regression/elastic net/lasso/ridge regression.

Conclusions: Models built using XGBoost and MLP are suitable for predicting readmission charges billed by hospitals. The MDCs allow models to accurately predict hospital readmission charges.

(JMIR Med Inform 2022;10(8):e37578) doi:10.2196/37578

KEYWORDS

readmission charges; readmission analytics; predictive models; machine learning; readmissions; predictive analytics

Introduction

Background

Electronic health records (EHRs) are now widely adopted by hospitals. EHR adoption has almost doubled since 2008, one of the reasons being the implementation of the government-related mandate as part of the American Recovery and Reinvestment Act of 2009 [1,2]. Even with the implementation of technological innovations like EHRs and various reforms for funding health care initiatives, health care costs have continued to increase. As per the recent National Health Expenditure Fact Sheet provided by the Centers for Medicare and Medicaid Services (CMS), the national health expenditure has grown 9.7% by the end of 2020, totaling US $4.1 trillion (approximately 19.7% of the Gross Domestic Product). On average, the United States of America spends over US $10,000 per resident per year toward health care. It is considerably higher than that in other countries included in the Organization for Economic Co-operation and Development, where the average cost is only US $4000 per person after adjusting for purchasing power [3].

Readmissions have been a significant contributor to rising health care costs. The hospital cost associated with 30-day all-cause readmissions was approximately US $41.3 billion for 2011 [4]. Even before the pandemic, annual hospital readmission costs were approximately US $26 billion for Medicare alone [5]. The pandemic caused a further increase in readmission costs [6]. Being expensive at the individual level, readmission is often postponed by patients until their health severely degenerates, leading to further increases in readmission costs, and these in turn contribute to the rapidly rising health care costs.

As a result, it is important for hospitals to plan for potential readmissions and associated costs. Although past research has primarily focused on predicting the probability of readmissions, the cost of readmissions is understudied, which is an important element in the financial planning done by hospitals as well as various concerned governmental agencies. As our task is to predict future hospital readmission charges, we take cues from existing literature on predictive analytics that have been applied and found beneficial in multiple areas of health care, such as risk analysis, disease diagnosis, disease progression, and preventive care [7-12]. Thus, we expect that predicting hospital readmission charges would help hospital policymakers plan for the upcoming expenditures. Hospitals can use these predictions to design policies based on the costs borne by individual patients.

According to the CMS, readmission is defined as an admission to a hospital within 30 days of discharge [13]. It could be from the same or another hospital, irrespective of the cause of readmission. However, readmission charges can be expected to vary significantly across major diagnostic categories (MDCs). To better control such variations and develop effective prediction models, we consider predicting the charges based on MDCs in this study, which is a novel aspect of our research. To the best of our knowledge, this aspect has not been explored in the past. We compare the predictions with the case when all diagnostic categories are pooled to predict readmission charges. Accordingly, we deploy the term readmission in two ways: readmission with the same major diagnostic category (RSDC) and all-cause readmission category (RADC). RSDC is defined as an admission to a hospital (same or another hospital) within 30 days of discharge with the cause of readmission being the same as the previous admission. In this context, the “cause of readmission” is based on the major diagnostic category (MDC). RADC is defined as an admission to a hospital within 30 days of readmission, irrespective of the cause of readmission.

Objective

The hospital charges for readmitted individuals can vary based on different services (such as procedures, labs, X-rays, and scans) used. Predicting these charges would be beneficial for financial planning by hospitals. Existing studies mainly focus on predicting either readmission probabilities or general health care costs. To date, no thorough research on suitable machine learning models exists for predicting hospital readmission charges. An exception is a study focusing on predicting readmission costs (not charges) [14]; however, it also does not include modern approaches, such as deep learning and regularization-based techniques. Our objective is to consider and compare traditional and modern predictive techniques to identify a suitable approach for predicting readmission charges.

Before building predictive models for RSDC and RADC, we also conducted preliminary analyses. First, for understanding the contribution of readmissions to the rising health care costs based on different criteria for readmissions (ie, RSDC and RADC), we determined the variation in the percentage of individuals contributing to hospital charges in our research context. Next, we analyzed whether readmissions varied across MDCs based on RSDC and RADC. As readmission policies vary across countries, we analyzed different readmission criteria for MDCs. Then, we determined whether the readmission charges changed significantly compared to the previous admission charges for RSDC and RADC. Finally, we strived to build models for predicting readmission charges billed by hospitals for RSDC and RADC.

Prior Work

The literature on applications of predictive methods for health care outcomes is vast. We focus on discussing works that directly relate to our study and context. Numerous machine learning–based approaches have been applied to predict readmissions and health care costs. For the sake of brevity, we list them succinctly in Table 1.

Table 1. Models used in prior studies.

Prediction area	Contexts and models used
Readmissions	All-cause: Artificial neural network (Jamei et al [15]); Bayesian network (Cai et al [16]); bidirectional encoder representation from transformers (Huang et al [17]); convolutional neural network (Wang et al [18]); Cox regression model (Yu et al [19]); decision trees (Sushmita et al [14] and Shadmi et al [20]); generalized boosting model (Sushmita et al [14]); multilayer perceptron (Wang et al [18]); multiple logistic regression (Sushmita et al [14], Schoonver et al [21], Picker et al [22], and Morris et al [23]); neural network (Shadmi et al [20] and Zheng et al [24]); random forest (Sushmita et al [14] and Zheng et al [24]); support vector machine (Sushmita et al [14], Yu et al [19], and Zheng et al [24]) Population-specific: Beta geometric Erlang-2 model (Bardhan et al [25]); lasso regularization with group-level feature selection (Radovanovic [26]); logistic regression (Yu et al [19], Kelly [27], and Hasan et al [28]); multivariate logistic regression (Tabata et al [29] and Greenblatt [30]); naïve Bayes (Shameer et al [31]); tree lasso logistic regression (Jovanovic et al [32]); multivariate Cox proportional hazard model (Schmutte et al [33]); XGBoost[a] (Morel et al [34])
Health care costs	General costs: Classification trees and clustering (Bertsimas et al [35]); linear regression (Farley et al [36], Sushmita et al [37], and Leigh et al [38]); M5 model tree (Sushmita et al [37]) High-cost patients: logistic regression (Fleishman and Cohen [39])

[a]XGBoost: eXtreme gradient boosting.

The first stream of research related to our study is on predicting readmissions. This body of literature is very large; therefore, we provide details on some representative research papers. A review paper [40] on readmission prediction models reports C statistic values between 0.55 and 0.65. Accordingly, the authors conclude that the models perform poorly. A recent study [41] reviewing articles from 2015 to 2019 reports an improvement in the C statistic values (greater than 0.75). For predicting readmissions, authors [21,42] explore the effects of physiological and medication regimens in some studies, whereas in another study [14], the authors use administrative data. Along these lines, existing studies [16,18,43] use machine learning approaches (such as deep learning and Bayesian network) to predict hospital readmission within 30 days. While using ensemble models, a model combining modified weighted boosting with a stacking algorithm shows a prediction performance 22% higher than that of a model combining the random forest algorithm, lasso algorithm, and Synthetic Minority Oversampling Technique [44]. A recent study [17] explores the use of unstructured data to predict readmission using bidirectional encoder representation from transformers. Extracting patient information from clinical notes using deep learning algorithms and then training them using graph neural networks is beneficial for prediction [45].

Next, focusing on specific subpopulation readmissions, past studies [25,29,31,46] use methods such as beta geometric Erlang-2, naïve Bayes, multivariate logistic regression, and tree-based lasso. In the case of readmissions with at least 7 past emergency department visits, boosted decision trees perform marginally better than logistic regression and the Bayes point machine [47]. A deep learning–based model built for congestive health failure patients using human-derived features, machine-derived contextual embeddings, and cost-sensitive sequential visit patterns in the EHR has the highest predictive power when compared to reduced models that use either 1 or more combinations of these [48]. eXtreme gradient boosting (XGBoost) shows better predictability than regularization techniques for predicting readmissions in mental or substance use disorders [34]. Interestingly, in a study related to psychiatric inpatients [33], the authors consider readmission within 12 months instead of the traditional 30 days to find which patient characteristics predict the time to readmission within 12 months. In terms of interpretability, existing studies [26,32] show that the tree-based lasso provides better interpretability. In an intensive care unit setting, attention-based networks may be preferable over recurrent neural networks when interpretability is of importance for a marginal decrease in accuracy [49]. Altogether, our literature review reveals that ensemble tree–based methods and deep learning approaches typically perform better than other approaches in predicting readmissions. However, none of the abovementioned studies predicts readmission charges, the focus of our study.

The literature closest to our work is on predicting health care–related costs. In one of the studies [35], the authors use classification trees and clustering algorithms to predict the general cost of health care and not specifically readmission charges. To apply these methods, the authors classify the continuous cost variable into discrete classes. In another study [50], the authors use more sophisticated machine learning methods, such as gradient boosting, an artificial neural network, and a ridge regression model, to predict cost-based classes. Although predicting general health care costs is useful, nothing can be concluded about the efficacy of these methods for predicting readmission charges because readmission is a fundamentally different phenomenon from general hospital visits. Specifically, readmission is usually associated with chronic illnesses and diseases requiring multiple visits. Moreover, bucketing a continuous variable into classes causes loss of information and may decrease predictive power.

There are studies [37,38] that predict general health care costs as a continuous variable. Apart from this, existing studies [51,52] derive costs based on predicting diagnosis-related groups (DRGs) to make operational decisions. However, as explained earlier, readmission charges are characteristically different from other types of costs. The prediction of readmission costs is considered in 1 study [14]. The authors use a limited set of methods, specifically linear regression and tree-based models, for predicting the costs (not charges). Based on our analysis of the existing literature, tree-based models and deep learning–based methods are likely to produce high prediction accuracies. We include a wide variety of prediction algorithms, including deep learning methods, to comprehensively study the

problem of predicting readmission charges. Moreover, we use a data set that spans the entire United States, unlike the existing study [14] that focuses on costs (not charges) using a data set with patients from a much smaller geographic region. Thus, we can provide robust recommendations on the methods that are best suited for making readmission charge predictions across different regions of the country.

Methods

Data Set and its Description

We used data from the Nationwide Readmission Database (NRD) of the Healthcare Cost and Utilization Project (HCUP) by the Agency for Healthcare Research and Quality for this study [53]. The data set consists of 4 parts, namely the core data set, severity data set, hospital-level data set, and diagnosis and procedure group data set. It includes inpatient individuals from the entire United States for 2013 (first fiscal year introducing readmission policies). Readmission policies and variables in the NRD data set have not changed much after that. We used nationally representative data available publicly to find generalizable insights that can be applied to all hospitals. The total number of records in the data set was 14,325,172, including those with and without repeat hospital visits. Initially, we analyzed readmissions with respect to hospital charges using the core data set part, which consists of hospital charges for an individual. We used variables from all 4 data set parts for building predictive models (see Multimedia Appendix 1 for the categorical and numeric variables used, along with their descriptive statistics and description). After cleaning the entire data set, we identified 576,701 and 1,091,580 individuals for the 2 readmission categories, namely RSDC and RADC, respectively. Each admission record consists of the following: demographics (gender, age, median household income, etc); clinical information (diagnosis, the procedure used, etc); comorbidities (hypertension, diabetes, depression, etc); hospital details (bed size, teaching or nonteaching hospital, etc); severity details (All Patients Refined Diagnosis Related Groups for severity of illness, risk of mortality, etc); and cost-related and administrative data (length of stay, charges billed by hospitals, etc).

The data set has close to 285 mutually exclusive categories of International Classification of Diseases (ICD-9) codes for grouping diagnoses and procedures related to patients for adjusting risks. Prior studies [26] have shown that aggregated higher-level grouping of diseases was effective in providing better results than going to a specific condition at the lowest level of hierarchy in the case of pediatric readmissions. MDC codes are at a higher level than the specific DRG payment codes in this context. Per the CMS, DRGs are grouped under MDCs formed focusing on a particular medical specialty and are mutually exclusive to make them clinically consistent. They are built based on principal diagnosis codes (ICD-9 codes in this data set).

We define the terms previous admission charge and average of previous admission charges used in this study. These terms differ for RSDC and RADC. The previous admission charge for RSDC is defined as the charge billed by the hospital for only the last previous admission having the same MDC. The previous admission charge for RADC is defined as the charge billed by the hospital for the last previous admission irrespective of the MDCs. The readmission charge for RSDC and RADC is defined as the charge billed by the hospital associated with 1 readmission visit using the readmission criteria based on the definitions of RSDC and RADC, respectively. The average of previous admission charges for RSDC is defined as the average charge billed by the hospital for all the previous admissions having the same MDC. The average of previous admission charges for RADC is defined as the average charge billed by the hospital for all the previous admissions, irrespective of the MDCs.

Ethical Considerations

We have signed the HCUP data use agreement. As per the HCUP data use agreement policy, HCUP databases are limited data sets. According to the Health Insurance Portability and Accountability Act of 1996 , review by an institutional review board is not required for limited data sets. Therefore, we did not apply for institutional review board approval for using the NRD data set [53].

Models Used and Their Description

The average previous admission charge was considered as one of the independent variables because the previous cost proved helpful in predicting future health care costs [35,37]. All the numeric independent variables were standardized except for the average admission charge for which log transformation was applied. Log transformation was also applied to the readmission charge, namely the dependent variable. We provide brief rationales behind the models considered for RSDC and RADC below.

Linear Regression (Baseline Model)

It is a simple and easily interpretable method compared to other nonlinear methods. It works well when there is a linear relationship between the dependent (target) variable and independent variables. We considered using linear regression as a baseline for this study, as it has been widely used for predicting general health care costs and is also computationally efficient [36-38].

Lasso Regression, Elastic Net Regression, and Ridge Regression

Regularization techniques prevent overfitting and multicollinearity by constraining the loss function. We could either add the penalty as the sum of the absolute values of coefficients (*L1* penalty) in lasso or as the sum of the squared values of coefficients in the case of ridge regression (*L2* penalty). Lasso gives us sparse solutions by shrinking the estimates for some coefficients to 0, whereas ridge regression shrinks the estimates near 0. Elastic net regression takes advantage of lasso and ridge regression by linearly combining *L1* and *L2* penalties. The literature review section explains that health-related data are complex and often face multicollinearity issues. To address these challenges, we applied regularization techniques to predict readmission charges billed by hospitals. In terms of hyperparameter tuning, we tuned α (that accounts for the relative importance of the lasso and ridge regression) ranging from 0 to 1 with a step size of 0.1, and estimated λ (the

regularization penalty) using cross-validation. The optimization objective in the hyperparameter tuning (in all methods used, including those introduced below) was set to minimize the root mean squared error (RMSE). We report results for lasso regression with α=1, elastic net regression with α=.5, and ridge regression with α=0.

XGBoost Model

It is one of the popular tree-based models for tabular data [54-56]. Prior studies [14] on predicting readmission costs (not charges) have also shown tree-based ensemble models to be beneficial. Therefore, we included this tree-based ensemble model for predicting readmission charges to take care of any nonlinearity. We chose XGBoost, as it has not been previously used in this context. Existing studies [57-59] show that a random search is sufficient and efficient in terms of the computation time for hyperparameter tuning. Hence, we performed a random search on the typical range of values for the relevant parameters depending on the type of booster [60]. The final values configured for this study are given in Table 2. The booster (type of learner) used was the tree booster (gbtree).

Table 2. eXtreme gradient boosting configuration details.

Configuration	Value
Number of rounds	120
Maximum depth of the tree	5
Learning rate	0.2
Subsample ratio of the training instances	0.7
L1[a] regularization term on weights	5
L2[b] regularization term on weights	20
Minimum loss required to make a split (gamma)	5
Subsample ratio of columns while constructing each tree	0.9

[a]L1: the sum of the absolute values of coefficients.

[b]L2: the sum of the squared values of coefficients.

Deep Learning Model Using Multilayer Perceptron

As discussed in the literature review section, even though deep learning–based models are more suitable for health-related data, there is no prior study that specifically predicts readmission charges using deep learning. A popular deep neural network architecture for tabular data is multilayer perceptron (MLP). Therefore, we used MLP, which requires multiple hyperparameters to be tuned. We chose the hyperparameters through a random search process, which is consistent with the recommendation provided in the literature pertaining to our case [57]. While choosing hyperparameters, we also used guidelines provided in relevant studies [61,62]. In our study, we found that models with even fewer hidden layers performed better than multiple linear regression. However, for the final configuration, we chose 4 hidden layers (beyond this, there was no further reduction in error values) to obtain fine-tuned low-error values and fewer epochs with consistent error values for the majority of the epochs. The values selected in this application are given in Table 3. The rectified linear unit was used as the activation function. The final activation function was linear, and the batch type was a minibatch.

Table 3. Configuration of the multilayer perceptron–based deep learning network.

Configuration	Value
Number of hidden layers	4
Number of neurons in the first hidden layer	80
Number of neurons in the second hidden layer	60
Number of neurons in the third hidden layer	50
Number of neurons in the fourth hidden layer	20
Minibatch size (weights get updated after each minibatch)	30
Momentum	0.9000
Learning rate	0.0001
Number of epochs (1 epoch = 1 forward pass + 1 backward pass)	400

Performance Measures Used

We used 7 metrics to measure the performance of the methods. We define n as the total number of observations (ie, patients), y_i as the actual values of readmission charges incurred by patients, $\bar{y}$ as the mean of readmission charges, and $\hat{y}$ as the

predicted values of readmission charges. The performance measures are provided below.

Mean Absolute Percentage Error

Mean absolute percentage error (MAPE) measures the error size in terms of percentage:

Root Mean Squared Error

Root mean squared error (RMSE) gives the standard deviation of the residual, which is the difference between actual and predicted values:

Mean Absolute Error

Mean absolute error (MAE) gives the average value of the errors for a given set of predictions:

Root Relative Squared Error

Root relative squared error (RRSE) gives the relative comparison of what the output would have been if a naïve model (simply predicting with the mean) were used:

Relative Absolute Error

Relative absolute error (RAE) compares the total absolute error of the model to the total absolute error of the simplest model (predicting with the mean):

Normalized Root Mean Squared Error

Normalized root mean squared error (NRMSE) is used to compare models with different scales:

$NRMSE = RMSE /$ 

Mean Absolute Deviation

Mean absolute deviation (MAD) describes how the values are spread away from the mean:

The lower the MAPE, RMSE, MAE, RAE, RRSE, NRMSE, and MAD, the better the prediction performance of the model.

Results

Initially, we analyzed the distribution of hospital charges (in percentage) contributed by individuals (in percentage) by giving different criteria for readmissions within RADC and RSDC, as shown in Figures 1 and 2. We found that 48% (US $294,802,405,683/US $614,171,678,507) of hospital charges came from 21% (2,108,143/10,038,776) of the individuals who had more than 1 admission.

Further analysis showed that the charges associated with readmissions varied from the initial admission charges for most diagnoses, with 541,090 individuals from the RADC category having readmission charges higher than the previous admission charges. Similarly, the current readmission charge was higher than the previous admission charge for 319,233 of the individuals for the RSDC category.

Next, we identified the MDCs having the highest number of readmissions for RADC and RSDC. The 2 groups are similar in terms of the MDCs with the highest number of readmissions. The categories with the highest number of readmissions for RSDC and RADC are given in Textbox 1 in descending order.

Next, we analyzed if the average readmission charge for each MDC in RSDC and RADC varied from the previous admission charge. In Figures 3 and 4, we explain the difference between the average readmission charge (ARC) and average previous admission charge (APAC) for RSDC and RADC, respectively. In the case of RSDC, the ARC was higher than the APAC for 21 of the MDCs (Figure 3). In contrast, the ARC was higher for only 13 of the MDCs in RADC (Figure 4).

We observed that readmission charges varied from previous admission charges at the individual and aggregated levels based on the above analysis. Next, we applied various predictive methods to predict readmission charges at an individual level for RSDC and RADC. We used 10-fold cross-validation. The test results are shown in Table 4 for RSDC and Table 5 for RADC.

Tables 4 and 5 show that the deep learning–based model and XGBoost performed the best compared to all the other models for all the performance metrics in RSDC and RADC. In addition, models such as lasso, elastic net, and ridge regression using regularization techniques on a linear model showed almost the same performance. Repeated measures ANOVA revealed that the mean RMSE differed significantly across models with $P<.001$. As ANOVA is an omnibus test, we also performed a post hoc test using the Bonferroni correction method. The test showed that the mean RMSE was statistically significantly ($P<.001$) lower for the deep learning/XGBoost models when compared to that of linear regression/elastic net/lasso/ridge regression. The test showed that the mean RMSE was statistically significantly ($P<.001$) lower for the deep learning and XGBoost models when compared to that of linear regression, elastic net, lasso, and ridge regression.

Figure 1. Distribution of hospital charges contributed by individuals (actual count in each category>10) for readmission with the same major diagnostic category.

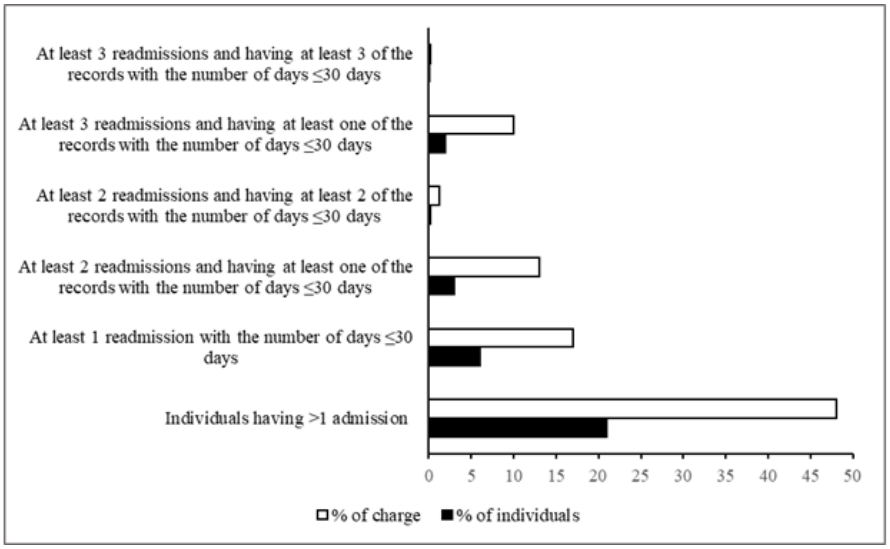
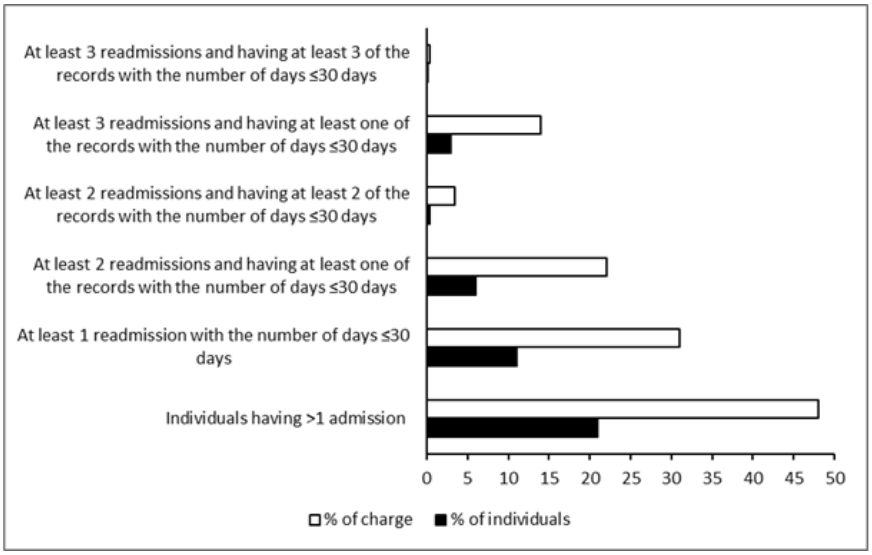


Figure 2. Distribution of hospital charges contributed by individuals (actual count in each category>10) for all-cause readmission category.



Textbox 1. Major diagnostic categories having the highest number of readmissions listed in descending order.

Readmission with the same major diagnostic category

- Diseases and disorders of the circulatory system
- Diseases and disorders of the respiratory system
- Diseases and disorders of the digestive system
- Infectious and parasitic diseases and disorders (systemic or unspecified sites)
- Diseases and disorders of the kidney and urinary tract
- Diseases and disorders of the nervous system

All-cause readmission category

- Diseases and disorders of the circulatory system
- Diseases and disorders of the respiratory system
- Diseases and disorders of the digestive system
- Pregnancy, childbirth, and puerperium
- Mental diseases and disorders
- Diseases and disorders of the nervous system

Figure 3. Difference between average readmission charge and average previous admission charge for readmission with the same major diagnostic category. MDC: major diagnostic category.

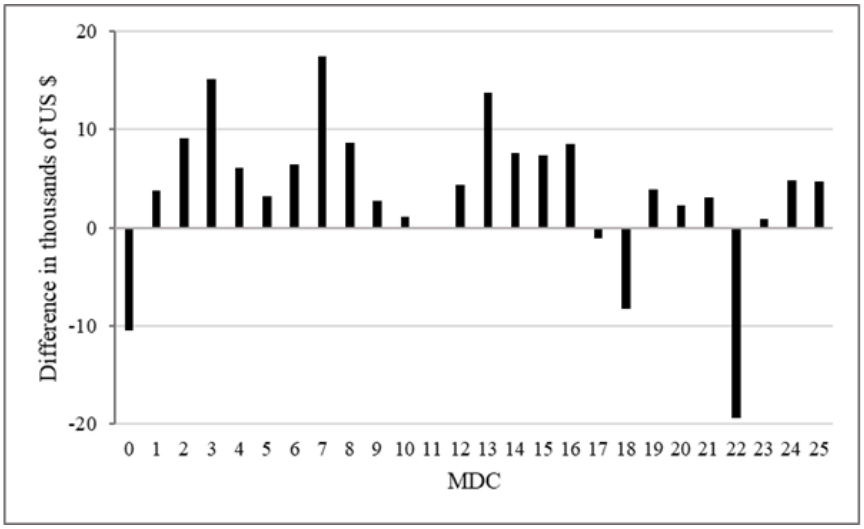


Figure 4. Difference between average readmission charge and average previous admission charge for all-cause readmission category. MDC: major diagnostic category.

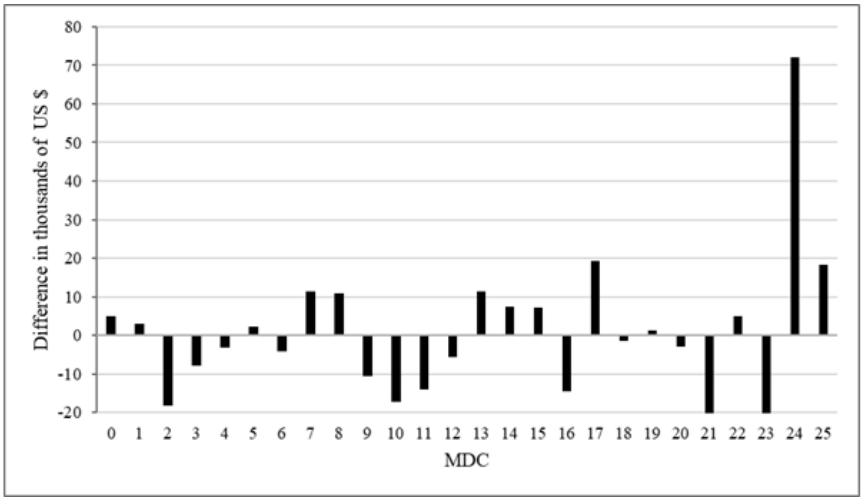


Table 4. Test results of readmission with the same major diagnostic category based on different performance measures.

Model	MAPE[a] (%), mean (SD)	RMSE[b], mean (SD)	MAE[c], mean (SD)	RRSE[d], mean (SD)	RAE[e], mean (SD)	NRMSE[f], mean (SD)	MAD[g], mean (SD)
Linear regression	4.268 (0.035)	0.564 (0.002)	0.431 (0.002)	0.546 (0.005)	0.528 (0.004)	0.055 (0.000)	0.042 (0.000)
Lasso	4.269 (0.036)	0.564 (0.002)	0.431 (0.002)	0.546 (0.005)	0.528 (0.004)	0.055 (0.000)	0.042 (0.000)
Elastic net	4.269 (0.036)	0.564 (0.002)	0.431 (0.002)	0.546 (0.005)	0.528 (0.004)	0.055 (0.000)	0.042 (0.000)
Ridge	4.299 (0.037)	0.565 (0.003)	0.434 (0.002)	0.547 (0.005)	0.531 (0.004)	0.055 (0.000)	0.042 (0.001)
XGBoost[h]	3.171 (0.027)	0.421 (0.003)	0.321 (0.002)	0.407 (0.004)	0.393 (0.003)	0.041 (0.001)	0.031 (0.000)
Deep learning	3.202 (0.022)	0.427 (0.003)	0.326 (0.002)	0.413 (0.004)	0.399 (0.003)	0.041 (0.001)	0.032 (0.000)

[a]MAPE: mean absolute percentage error.

[b]RMSE: root mean squared error.

[c]MAE: mean absolute error.

[d]RRSE: root relative squared error.

[e]RAE: relative absolute error.

[f]NRMSE: normalized root mean squared error.

[g]MAD: mean absolute deviation.

[h]XGBoost: eXtreme gradient boosting.

Table 5. Test results of all-cause readmission category based on different performance measures.

Model	MAPE[a] (%), mean (SD)	RMSE[b], mean (SD)	MAE[c], mean (SD)	RRSE[d], mean (SD)	RAE[e], mean (SD)	NRMSE[f], mean (SD)	MAD[g], mean (SD)
Linear regression	4.208 (0.047)	0.558 (0.004)	0.427 (0.003)	0.554 (0.005)	0.537 (0.005)	0.054 (0.000)	0.041 (0.001)
Lasso	4.208 (0.047)	0.558 (0.004)	0.427 (0.003)	0.554 (0.005)	0.537 (0.005)	0.054 (0.000)	0.041 (0.001)
Elastic net	4.209 (0.047)	0.558 (0.004)	0.427 (0.003)	0.554 (0.005)	0.537 (0.005)	0.054 (0.000)	0.041 (0.001)
Ridge	4.240 (0.049)	0.559 (0.005)	0.429 (0.003)	0.555 (0.005)	0.531 (0.005)	0.054 (0.000)	0.042 (0.001)
XGBoost[h]	3.121 (0.019)	0.414 (0.002)	0.317 (0.002)	0.410 (0.001)	0.399 (0.002)	0.040 (0.000)	0.031 (0.000)
Deep learning	3.103 (0.018)	0.413 (0.003)	0.316 (0.003)	0.410 (0.002)	0.397 (0.003)	0.040 (0.000)	0.031 (0.000)

[a]MAPE: mean absolute percentage error.

[b]RMSE: root mean squared error.

[c]MAE: mean absolute error.

[d]RRSE: root relative squared error.

[e]RAE: relative absolute error.

[f]NRMSE: normalized root mean squared error.

[g]MAD: mean absolute deviation.

[h]XGBoost: eXtreme gradient boosting.

Discussion

Principal Results and Comparison With Prior Work

This study shows that national administrative data can be used to build effective predictive models for hospital charges billed for readmissions, even if there are different criteria for readmissions. The deep learning–based algorithm and XGBoost outperformed all other algorithms. Based on our experiments, we also made a few observations specific to configuring XGBoost. While configuring the XGBoost model, we found that using the gradient descent of the tree-type booster gave the best performance compared to other boosters such as a linear booster or dropouts with multiple additive regression tree boosters. Moreover, in the same context, setting the booster to linear with regularization for XGBoost gave a performance equivalent to linear, lasso, elastic net, and ridge regression.

In summary, this study makes 2 important contributions. To the best of our knowledge, this is the first study to apply regularization techniques, a tree-based ensemble model using XGBoost, and deep learning–based models for predicting readmission charges billed by hospitals. Deep learning–based models and XGBoost have proven useful in modeling health-related data. A related study that focused on predicting readmission costs (not charges) used only linear regression and tree-based models on narrow data sets (~10k samples) with limited features, and hence, its applicability in different geographies is questionable. Besides, it predicted readmission costs (not charges) using only the all-cause definition of readmission. Our study considered readmission using MDCs

instead of DRGs by using different MDC criteria to determine which models would be suitable for predicting readmission charges.

Implications

This study has 2 practical implications. First, health systems use high-risk care management programs to improve health outcomes in individuals with complex needs and reduce costs. As these programs are resource-intensive and expensive, health systems use costs as a proxy to identify individuals suitable for these programs [63]. Our study related to readmissions will aid such programs by prescribing models that will provide reliable estimates of readmission charges.

Second, hospital reimbursement mainly depends on DRG codes and the case mix index (CMI). The CMI is calculated as the average DRG weight of the hospitals' inpatient discharges. A higher CMI would indicate more reimbursement for hospitals. As the CMI is not directly tied to either hospital charges (which can vary depending on various factors specific to the hospital, such as staffing expenses and technologies used) or individual-specific expenses, hospitals often do not get reimbursed for the services they have provided [64]. In this study, we predicted readmission charges that will give hospitals a better estimate of the cost they are going to incur in case the patients get readmitted. Now, hospitals can use the CMI and DRGs to determine their reimbursement amounts and compare that with the estimated charges. If there are any differences in the amount, hospitals can now more effectively plan for mitigation strategies. Thus, in a nutshell, our study can be helpful for health care policymakers and hospital planners.

Limitations and Future Research

Modeling readmission likelihood and the length of stay are also crucial in readmissions, as these outcomes influence one another. Moreover, modeling readmission charges, readmission likelihood, and length of stay might be more beneficial than focusing only on modeling readmission charges. In this study, we identified readmissions belonging to RSDC and RADC. We will also use the term readmission in the readmission with different major diagnostic category (RDDC) for our future analysis. RDDC will consider readmission as an admission to a hospital within 30 days of discharge from the same or another hospital with the cause of readmission being different. We will then build predictive models for RDDC. Then, we will compare the predictive models built for RDDC with those built for RSDC.

In this study, we considered the standard defined categories of MDCs as the cause of readmission. The standard defined categories of MDCs belong to either a single organ system or an etiology. For our future study, we will consider correlated categories in terms of the set of related health complications that eventually lead to readmissions. These categories may span multiple MDCs. We expect that such recategorizations could help in the better prediction of charges. The recategorization in terms of correlated categories would significantly contribute to health care economics.

Conclusions

Readmissions are one of the main contributors to health care costs. However, most previous studies have focused mainly on predicting early readmissions. The implementation of the Hospital Readmissions Reduction Program has mixed reviews, with no conclusion regarding its effectiveness. This study aimed to determine if readmission charges, which vary from initial admission charges, could be accurately predicted. Results revealed that the deep learning–based model and XGBoost performed the best in terms of all performance measures. MDCs can be used to accurately predict charges billed by hospitals for readmissions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variables used in this study, along with their descriptions and descriptive statistics.
[PDF File (Adobe PDF File), 181 KB - medinform_v10i8e37578_app1.pdf]

References

1. Jamoom E, Yang N. Table of electronic health record adoption use among office-based physicians in the US by state: 2015 National Electronic Health Records Survey. National Electronic Health Records Survey: 2015 State and National Electronic Health Record Adoption Summary Tables. Hyattsville, MD, United States: National Center for Health Statistics; 2016. URL: https://www.cdc.gov/nchs/data/ahcd/nehrs/2015_nehrs_web_table.pdf [accessed 2021-12-29]
2. Atherton J. Development of the electronic health record. AMA J Ethics 2011 Mar;13(3):186-189 [FREE Full text] [doi: 10.1001/virtualmentor.2011.13.3.mhst1-1103] [Medline: 23127323]
3. OECD. Health at a Glance 2019: OECD Indicators. Paris, France: OECD Publishing; 2019.
4. Hines AL, Barrett ML, Jiang HJ, Steinar CA. Conditions with the largest number of adult hospital readmissions by payer, 2011. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Brief #172. Rockville, MD, United States: Agency for Healthcare Research and Quality; Apr 2014.
5. Wilson L. MA patients' readmission rates higher than traditional Medicare, study finds; 2019. HEALTHCAREDIVE. 2019 Jun. URL: https://www.healthcaredive.com/news/ma-patients-readmission-rates-higher-than-traditional-medicare-study-find/557694/ [accessed 2021-01-11]

6. LaPointe J. CDC: 1 in 11 COVID-19 inpatients experience a hospital readmission. Xtelligent Healthcare Media. 2020 Nov. URL: https://revcycleintelligence.com/news/cdc-1-in-11-covid-19-inpatients-experience-a-hospital-readmission [accessed 2021-01-11]
7. Vaid A, Jaladanki S, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. JMIR Med Inform 2021 Jan;9(1):e24207 [FREE Full text] [doi: 10.2196/24207] [Medline: 33400679]
8. Tran L, Chi L, Bonti A, Abdelrazek M, Phoebe Chen YP. Mortality prediction of patients with cardiovascular disease using medical claims data under artificial intelligence architectures: validation study. JMIR Med Inform 2021 Apr;9(4):e25000 [FREE Full text] [doi: 10.2196/25000] [Medline: 33792549]
9. Zhao P, Yoo I, Naqvi SH. Early prediction of unplanned 30-day hospital readmission: model development and retrospective data analysis. JMIR Med Inform 2021 Mar;9(3):e16306 [FREE Full text] [doi: 10.2196/16306] [Medline: 33755027]
10. Conway A, Jungquist CR, Chang K, Kamboj N, Sutherland J, Mafeld S, et al. Predicting prolonged apnea during nurse-administered procedural sedation: machine learning study. JMIR Perioper Med 2021 Oct;4(2):e29200 [FREE Full text] [doi: 10.2196/29200] [Medline: 34609322]
11. Hou C, Zhong X, He P, Xu B, Diao S, Yi F, et al. Predicting breast cancer in Chinese women using machine learning techniques: algorithm development. JMIR Med Inform 2020 Jun;8(6):e17364 [FREE Full text] [doi: 10.2196/17364] [Medline: 32510459]
12. Lee E, Jung SY, Hwang HJ, Jung J. Patient-level cancer prediction models from a nationwide patient cohort: model development and validation. JMIR Med Inform 2021 Aug;9(8):e29807 [FREE Full text] [doi: 10.2196/29807] [Medline: 34459743]
13. Payment policy for inpatient readmissions promoting greater efficiency in Medicare. MedPAC. 2007. URL: https://www.medpac.gov/wp-content/uploads/import_data/scrape_files/docs/default-source/reports/Jun07_Ch05.pdf [accessed 2022-08-04]
14. Sushmita S, Khulbe G, Hasan A, Newman S, Ravindra P, Roy SB, et al. Predicting 30-day risk and cost of "all-cause" hospital readmission. In: The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence Expanding the Boundaries of Health Informatics Using AI: Technical Report WS-16-08. 2016 Presented at: 30th AAAI Conference on Artificial Intelligence; February 12-13, 2016; Phoenix, AZ.
15. Jamei M, Nisnevich A, Wetchler E, Sudat S, Liu E. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PLoS ONE 2017 Jul;12(7):e0181173. [doi: 10.1371/journal.pone.0181173] [Medline: 28708848]
16. Cai X, Perez-Concha O, Coiera E, Martin-Sanchez F, Day R, Roffe D, et al. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. J Am Med Inform Assoc 2016 May;23(3):553-561 [FREE Full text] [doi: 10.1093/jamia/ocv110] [Medline: 26374704]
17. Huang K, Altosaar J, Ranganathan R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. ArXiv Preprint posted online Apr 10, 2019. [doi: 10.48550/arXiv.1904.05342]
18. Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting hospital readmission via cost-sensitive deep learning. IEEE/ACM Trans Comput Biol Bioinform 2018 Nov;15(6):1968-1978. [doi: 10.1109/TCBB.2018.2827029] [Medline: 29993930]
19. Yu S, Farooq F, van Esbroeck A, Fung G, Anand V, Krishnapuram B. Predicting readmission risk with institution-specific prediction models. Artif Intell Med 2015 Oct;65(2):89-96. [doi: 10.1016/j.artmed.2015.08.005] [Medline: 26363683]
20. Shadmi E, Flaks-Manov N, Hoshen M, Goldman O, Bitterman H, Balicer RD. Predicting 30-day readmissions with preadmission electronic health record data. Med Care 2015 Mar;53(3):283-289. [doi: 10.1097/MLR.0000000000000315] [Medline: 25634089]
21. Schoonover H, Corbett CF, Weeks DL, Willson MN, Setter SM. Predicting potential postdischarge adverse drug events and 30-day unplanned hospital readmissions from medication regimen complexity. J Patient Saf 2014 Dec;10(4):186-191. [doi: 10.1097/PTS.0000000000000067] [Medline: 25408236]
22. Picker D, Heard K, Bailey TC, Martin NR, LaRossa GN, Kollef MH. The number of discharge medications predicts thirty-day hospital readmission: a cohort study. BMC Health Serv Res 2015 Jul;15:282 [FREE Full text] [doi: 10.1186/s12913-015-0950-9] [Medline: 26202163]
23. Morris PE, Griffin L, Berry M, Thompson C, Hite RD, Winkelman C, et al. Receiving early mobility during an intensive care unit admission is a predictor of improved outcomes in acute respiratory failure. Am J Med Sci 2011 May;341(5):373-377 [FREE Full text] [doi: 10.1097/MAJ.0b013e31820ab4f6] [Medline: 21358312]
24. Zheng B, Zhang J, Yoon SW, Lam SS, Khasawneh M, Poranki S. Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Syst Appl 2015 Nov;42(20):7110-7120. [doi: 10.1016/j.eswa.2015.04.066]
25. Bardhan I, Jeong-ha (, Oh Z(, Kirk, Kirksey Z. Predictive analytics for readmission of patients with congestive heart failure. Inf Syst Res 2015 Mar;26(1):19-39. [doi: 10.1287/isre.2014.0553]
26. Radovanovic S, Vukicevic M, Kovacevic A, Stiglic G, Obradovic Z. Domain knowledge based hierarchical feature selection for 30-day hospital readmission prediction. In: Holmes J, Bellazzi R, Sacchi L, Peek N, editors. Artificial Intelligence in Medicine. AIME 2015. Lecture Notes in Computer Science(), vol 9105. Cham, Switzerland: Springer; Jun 2015:96-100.

27. Kelly M, Sharp L, Dwane F, Kelleher T, Comber H. Factors predicting hospital length-of-stay and readmission after colorectal resection: a population-based study of elective and emergency admissions. BMC Health Serv Res 2012 Mar 26;12:77 [FREE Full text] [doi: 10.1186/1472-6963-12-77] [Medline: 22448728]
28. Hasan O, Meltzer DO, Shaykevich SA, Bell CM, Kaboli PJ, Auerbach AD, et al. Hospital readmission in general medicine patients: a prediction model. J Gen Intern Med 2010 Mar;25(3):211-219 [FREE Full text] [doi: 10.1007/s11606-009-1196-1] [Medline: 20013068]
29. Tabata M, Shimizu R, Kamekawa D, Kato M, Kamiya K, Akiyama A, et al. Six-minute walk distance is an independent predictor of hospital readmission in patients with chronic heart failure. Int Heart J 2014 Oct;55(4):331-336 [FREE Full text] [doi: 10.1536/ihj.13-224] [Medline: 24898596]
30. Greenblatt DY, Weber SM, O'Connor ES, LoConte NK, Liou J, Smith MA. Readmission after colectomy for cancer predicts one-year mortality. Ann Surg 2010 Apr;251(4):659-669 [FREE Full text] [doi: 10.1097/SLA.0b013e3181d3d27c] [Medline: 20224370]
31. Shameer K, Johnson KW, Yahi A, Miotto R, Li LI, Ricks D, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai heart failure cohort. Pac Symp Biocomput 2017;22:276-287 [FREE Full text] [doi: 10.1142/9789813207813_0027] [Medline: 27896982]
32. Jovanovic M, Radovanovic S, Vukicevic M, Van Poucke S, Delibasic B. Building interpretable predictive models for pediatric hospital readmission using tree-lasso logistic regression. Artif Intell Med 2016 Sep;72:12-21. [doi: 10.1016/j.artmed.2016.07.003] [Medline: 27664505]
33. Schmutte T, Dunn CL, Sledge WH. Predicting time to readmission in patients with recent histories of recurrent psychiatric hospitalization: a matched-control survival analysis. J Nerv Ment Dis 2010 Dec;198(12):860-863. [doi: 10.1097/NMD.0b013e3181fe726b] [Medline: 21135635]
34. Morel D, Yu KC, Liu-Ferrara A, Caceres-Suriel AJ, Kurtz SG, Tabak YP. Predicting hospital readmission in patients with mental or substance use disorders: a machine learning approach. Int J Med Inform 2020 Jul;139:104136 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104136] [Medline: 32353752]
35. Bertsimas D, Bjarnadóttir MV, Kane MA, Kryder JC, Pandey R, Vempala S, et al. Algorithmic prediction of health-care costs. Oper Res 2008 Dec;56(6):1382-1392. [doi: 10.1287/opre.1080.0619]
36. Farley JF, Harley CR, Devine JW. A comparison of comorbidity measurements to predict healthcare expenditures. Am J Manag Care 2006 Feb;12(2):110-119 [FREE Full text] [Medline: 16464140]
37. Sushmita S, Newman S, Marquardt J, Ram P, Prasad V, Cock MD, et al. Population cost prediction on public healthcare datasets. In: DH '15: Proceedings of the 5th International Conference on Digital Health 2015. New York, NY, United States: Association for Computing Machinery; 2015 May Presented at: DH '15: Digital Health 2015 Conference; May 18-20, 2015; Florence, Italy p. 87-94. [doi: 10.1145/2750511.2750521]
38. Leigh JP, Hubert HB, Romano PS. Lifestyle risk factors predict healthcare costs in an aging cohort. Am J Prev Med 2005 Dec;29(5):379-387. [doi: 10.1016/j.amepre.2005.08.005] [Medline: 16376700]
39. Fleishman JA, Cohen JW. Using information on clinical conditions to predict high-cost patients. Health Serv Res 2010 Apr;45(2):532-552 [FREE Full text] [doi: 10.1111/j.1475-6773.2009.01080.x] [Medline: 20132341]
40. 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;306(15):1688-1698 [FREE Full text] [doi: 10.1001/jama.2011.1515] [Medline: 22009101]
41. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ 2020 Apr;369:m958 [FREE Full text] [doi: 10.1136/bmj.m958] [Medline: 32269037]
42. Xue Y, Klabjan D, Luo Y. Predicting ICU readmission using grouped physiological and medication trends. Artif Intell Med 2019 Apr;95:27-37 [FREE Full text] [doi: 10.1016/j.artmed.2018.08.004] [Medline: 30213670]
43. Xiao C, Ma T, Dieng AB, Blei DM, Wang F. Readmission prediction via deep contextual embedding of clinical concepts. PLoS One 2018 Apr;13(4):e0195024 [FREE Full text] [doi: 10.1371/journal.pone.0195024] [Medline: 29630604]
44. Yu K, Xie X. Predicting hospital readmission: a joint ensemble-learning model. IEEE J Biomed Health Inform 2020 Feb;24(2):447-456. [doi: 10.1109/JBHI.2019.2938995] [Medline: 31484143]
45. Golmaei SN, Luo X. DeepNote-GNN: predicting hospital readmission using clinical notes and patient network. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. New York, NY, United States: Association for Computing Machinery; 2021 Presented at: 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; August 1-4, 2021; Gainesville, Florida, United States URL: https://doi.org/10.1145/3459930.3469547 [doi: 10.1145/3459930.3469547]
46. Cui S, Wang D, Wang Y, Yu P, Jin Y. An improved support vector machine-based diabetic readmission prediction. Comput Methods Programs Biomed 2018 Nov;166:123-135. [doi: 10.1016/j.cmpb.2018.10.012] [Medline: 30415712]
47. Ben-Assuli O, Padman R. Analysing repeated hospital readmissions using data mining techniques. Health Syst (Basingstoke) 2018 Nov;7(3):166-180 [FREE Full text] [doi: 10.1080/20476965.2018.1510040] [Medline: 31215903]
48. Ashfaq A, Sant'Anna A, Lingman M, Nowaczyk S. Readmission prediction using deep learning on electronic health records. J Biomed Inform 2019 Sep;97:103256 [FREE Full text] [doi: 10.1016/j.jbi.2019.103256] [Medline: 31351136]

49. Barbieri S, Kemp J, Perez-Concha O, Kotwal S, Gallagher M, Ritchie A, et al. Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk. Sci Rep 2020 Jan;10:1111 [FREE Full text] [doi: 10.1038/s41598-020-58053-z] [Medline: 31980704]
50. Morid MA, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. AMIA Annu Symp Proc 2017 Apr;2017:1312-1321 [FREE Full text] [Medline: 29854200]
51. Liu J, Capurro D, Nguyen A, Verspoor K. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. NPJ Digit Med 2021 Jul;4:103 [FREE Full text] [doi: 10.1038/s41746-021-00474-9] [Medline: 34211109]
52. Gartner D, Kolisch R, Neill DB, Padman R. Machine learning approaches for early DRG classification and resource allocation. INFORMS J Comput 2015 Nov;27(4):718-734 [FREE Full text] [doi: 10.1287/ijoc.2015.0655]
53. HCUP Nationwide Readmissions Database (NRD). Healthcare Cost Utilization Project (HCUP). Agency for Healthcare Research Quality. 2013. URL: https://www.hcup-us.ahrq.gov/nrdoverview.jsp [accessed 2021-01-11]
54. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion 2022 May;81:84-90 [FREE Full text] [doi: 10.1016/j.inffus.2021.11.011]
55. Edgari E, Thiojaya J, Qomariyah NN. The impact of Twitter sentiment analysis on bitcoin price during COVID-19 with XGBoost. : IEEE; 2022 Presented at: 2022 5th International Conference on Computing and Informatics (ICCI); March 9-10, 2022; New Cairo, Egypt. [doi: 10.1109/ICCI54321.2022.9756123]
56. Sahin EK. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Appl Sci 2020 Jun;2(7):1-17 [FREE Full text] [doi: 10.1007/s42452-020-3060-1]
57. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13(2012):281-305 [FREE Full text]
58. Mantovani RG, Rossi AL, Vanschoren J, Bischl B, De Carvalho AC. Effectiveness of random search in SVM hyper-parameter tuning. : IEEE; 2015 Jul Presented at: International Joint Conference on Neural Networks (IJCNN); July 12-17, 2015; Killarney, Ireland p. 1-8. [doi: 10.1109/ijcnn.2015.7280664]
59. Park J, Lee Y, Lee J. Assessment of machine learning algorithms for land cover classification using remotely sensed data. Sens Mater 2021 Nov;33(11):3885-3902 [FREE Full text] [doi: 10.18494/sam.2021.3612]
60. Wade C. Hands-On Gradient Boosting With XGBoost and scikit-learn: Perform Accessible Machine Learning and Extreme Gradient Boosting With Python. Birmingham, United Kingdom: Packt Publishing Ltd; 2020.
61. Bengio Y. Practical recommendations for gradient-based training for deep architectures. In: Neural networks: Tricks of the trade. New York City, NY: Springer; 2012:437-478.
62. Sheela K, Deepa SN. Review on methods to fix number of hidden neurons in neural networks. Math Probl Eng 2013;2013:1-11 [FREE Full text] [doi: 10.1155/2013/425740]
63. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
64. Ericson C. Is case mix index still a relevant key performance indicator? Journal of AHIMA. 2021 Jan. URL: https://journal.ahima.org/page/is-case-mix-index-still-a-relevant-key-performance-indicator [accessed 2022-06-20]

Abbreviations

APAC: average previous admission charge
ARC: average readmission charge
CMI: case mix index
CMS: Centers for Medicare and Medicaid Services
DRG: diagnosis-related group
EHR: electronic health record
HCUP: Healthcare Cost and Utilization Project
ICD: International Classification of Diseases
MAD: mean absolute deviation
MAE: mean absolute error
MAPE: mean absolute percentage error
MDC: major diagnostic category
MLP: multilayer perceptron
NRD: Nationwide Readmission Database
NRMSE: normalized root mean squared error
RADC: all-cause readmission category
RAE: relative absolute error
RDDC: readmission with different major diagnostic category

RMSE: root mean squared error
RRSE: root relative squared error
RSDC: readmission with the same major diagnostic category
XGBoost: eXtreme gradient boosting

Edited by C Lovis, J Hefner; submitted 25.02.22; peer-reviewed by D Gartner; comments to author 18.03.22; revised version received 02.05.22; accepted 26.07.22; published 30.08.22.

Please cite as:
Gopukumar D, Ghoshal A, Zhao H
Predicting Readmission Charges Billed by Hospitals: Machine Learning Approach
JMIR Med Inform 2022;10(8):e37578
URL: https://medinform.jmir.org/2022/8/e37578
doi:10.2196/37578
PMID:35896038

Original Paper

A Machine Learning Approach for Continuous Mining of Nonidentifiable Smartphone Data to Create a Novel Digital Biomarker Detecting Generalized Anxiety Disorder: Prospective Cohort Study

Soumya Choudhary[1], BSc, MSc; Nikita Thomas[2], BSc, MSc; Sultan Alshamrani[2], BSc, MSc, PhD; Girish Srinivasan[2], BSc, MSc, PhD; Janine Ellenberger[1], MD; Usman Nawaz[2], BSc, MSc; Roy Cohen[1], BSc, MSc

[1]Department of Research, Behavidence, Inc., New York, NY, United States

[2]Department of Data Science, Behavidence, Inc., New York, NY, United States

Corresponding Author:
Soumya Choudhary, BSc, MSc
Department of Research
Behavidence, Inc.
99 Wall Street, Suite Number 4004
New York, NY, 10005
United States
Phone: 1 8477228324
Email: soumya@behavidence.com

Abstract

Background: Anxiety is one of the leading causes of mental health disability around the world. Currently, a majority of the population who experience anxiety go undiagnosed or untreated. New and innovative ways of diagnosing and monitoring anxiety have emerged using smartphone sensor–based monitoring as a metric for the management of anxiety. This is a novel study as it adds to the field of research through the use of nonidentifiable smartphone usage to help detect and monitor anxiety remotely and in a continuous and passive manner.

Objective: This study aims to evaluate the accuracy of a novel mental behavioral profiling metric derived from smartphone usage for the identification and tracking of generalized anxiety disorder (GAD).

Methods: Smartphone data and self-reported 7-item GAD anxiety assessments were collected from 229 participants using an Android operating system smartphone in an observational study over an average of 14 days (SD 29.8). A total of 34 features were mined to be constructed as a potential digital phenotyping marker from continuous smartphone usage data. We further analyzed the correlation of these digital behavioral markers against each item of the 7-item Generalized Anxiety Disorder Scale (GAD-7) and its influence on the predictions of machine learning algorithms.

Results: A total of 229 participants were recruited in this study who had completed the GAD-7 assessment and had at least one set of passive digital data collected within a 24-hour period. The mean GAD-7 score was 11.8 (SD 5.7). Regression modeling was tested against classification modeling and the highest prediction accuracy was achieved from a binary XGBoost classification model (precision of 73%-81%; recall of 68%-87%; F_1-score of 71%-79%; accuracy of 76%; area under the curve of 80%). Nonparametric permutation testing with Pearson correlation results indicated that the proposed metric (Mental Health Similarity Score [MHSS]) had a colinear relationship between GAD-7 Items 1, 3 and 7.

Conclusions: The proposed MHSS metric demonstrates the feasibility of using passively collected nonintrusive smartphone data and machine learning–based data mining techniques to track an individuals' daily anxiety levels with a 76% accuracy that directly relates to the GAD-7 scale.

(JMIR Med Inform 2022;10(8):e38943) doi:10.2196/38943

KEYWORDS

digital phenotyping; machine learning; mental health; profiling metric; smartphone data; anxiety assessment; mining technique; algorithm prediction; digital marker; behavioral marker; anxiety

Introduction

Background and Rationale

Anxiety is one of the leading causes of mental health disability around the world [1]. It includes feelings of excessive worry and negative thoughts, accompanied by physical symptoms such as heart palpitations and increased blood pressure [2]. Anxiety is also associated with a high degree of functional impairment [3] leading to poor quality of life [4] and high health care utilization [5]. Despite being one of the leading causes of mental health disability (1 in 4 people according to the World Mental Health Survey [6]), the detection of generalized anxiety disorder (GAD) is very low in primary care settings [7-9]. These challenges stem from the problems regarding diagnostic processes and inaccuracies [8,10-16] as well as overlapping comorbidities [9,17,18] and physical symptomatology [5,19]. The diagnosis is also vulnerable to the observer's state of mind [20] and biased self-perception [21] of symptoms. Whether it is the diagnosis of GAD as a singular condition or as a comorbidity, the validity of the diagnostic classifications and instruments in themselves has been rigorously debated. Newson et al [22] highlighted the heterogeneity in DSM-5 classification, where it failed to diagnose a specific disorder from random. Zimmerman et al [23] demonstrated how a physician can diagnose depression and its comorbidities in 227 different ways and Phillips [15] has highlighted the ambiguities in DSM-5 criteria for disorder classification. A recent analysis [10] of eHealth data, patient records, and physician reports in psychiatric cases has highlighted the presence of diagnostic errors in two-thirds of the sample.

With the advancement of technology, researchers have employed multisource data and advanced data analysis techniques to refine and improve mental health diagnosis. One such opportunity to use an upcoming method to improve screening of anxiety is to harness the power of smartphones using the principles of digital phenotyping [24]. Digital phenotyping is a novel computational approach that relies on real-time quantification of human behavior through continuous monitoring of digital biomarkers [25-27]. Mobile and wearable digital devices offer the opportunity to track a multitude of parameters such as mobility (through GPS and accelerometer) [28,29], societal interactions [30] (number of calls, voice tone detection, number of messages sent), digital interactions (access to certain apps), phone usage frequency (screen turned on/off) [27], and health monitoring parameters (heart rate, blood pressure, and oxygen saturation) [31]. However, most digital phenotyping approaches present limited applicability due to the lack of standardized data processing approach for big data exploitation and lack of a specific pattern of unique features for complex mental conditions such as anxiety disorder.

Previous Findings

Smartphones hold huge potential in redefining the ability to understand mental health behavior. Sensors embedded in smartphones allow for both passive and continuous data collection, which enhances the possibility of understanding human behavior daily [32-34]. Longitudinal monitoring of passive sensors and phone usage has been linked to tracking mental health behavioral trends [24]. Digital phenotyping of mental health has proven successful in dealing with the challenges associated with a diagnosis such as biases in self-reporting and lack of time in primary care settings, thus paving the way for new and novel methods of screening and monitoring [35].

Most previous studies have focused on using digital phenotyping and passive sensor data to predict social anxiety rather than generalized anxiety [28,29,32,36]. In addition, the passive data used in previous research were intrusive of the users' privacy and collected identifiable data points such as GPS, audio, message logs, and Bluetooth. Jacobson et al [29] demonstrated that sensor data such as accelerometer, call log, and text message data from smartphones could predict social anxiety symptom severity. Another study found that people with high social anxiety had much lower call and text message logs, and used more health and fitness apps and less camera apps as compared with the low social anxiety group [36]. A clinical review on digital phenotyping and the mental health of college students found that sensors such as accelerometer, Bluetooth, and social information can help in understanding clinical symptomatology [37]. By contrast, Meyerhoff et al [28] found that GPS-based sensor features can be useful in predicting depression severity, but it was not significant in predicting anxiety. Other studies that have researched generalized anxiety have been grouped along with other disorders such as depression and social anxiety. The sensors that have been utilized included location sampled every 5 minutes, call and message log data, duration, and length. Interestingly, these studies also found that there was no significant relationship between GAD and location sensors [28,38]. A more recent study investigated how features extracted from smartphones can be used to predict GAD, social anxiety disorder, and depression. The authors found that their machine learning models and features were able to predict social anxiety disorder and depression severity but not GAD [25]. Such findings have paved the way to explore more ways to map generalized anxiety using nonintrusive and nonidentifiable smartphone data.

Study Objective

In this study a novel mental behavioral profiling metric, derived from smartphone usage, is defined for the identification and tracking of GAD. The accuracy of this metric is evaluated in relation to the standardized anxiety assessment protocol using the 7-item Generalized Anxiety Disorder Scale (GAD-7) questionnaire scoring.

Methods

Data Collection Procedure

Participants were recruited via an advertisement through social media campaigns on Facebook and Google. Research has shown that this is an effective means of recruitment and provides more generalizability than a clinic-recruited study [39]. Interested participants responded to the advertisement by reading about the study and signing the informed consent form. They then downloaded the "Behavidence Research App" from the Google Play store and filled in a demographic questionnaire, followed by the GAD-7 scale. These data were collected at a single time

point only during the onboarding process. The app continued to passively collect nonintrusive data from the smartphone such as screen time and app usage, with no engagement requirement from the user. There was absolutely no private information collected, making this solution completely nonintrusive and secure. Data were collected between October 2021 and January 2022. The participants were informed about the type of nonidentifiable passive data collected in the consent form.

Inclusion/Exclusion Criteria

A total of 238 globally distributed users responded to the online advertisement. The inclusion criteria were (1) participants should be over 18 years of age; (2) participants must be able read, speak, and write in English; and (3) participants must have an Android smartphone. Of the enrolled participants, 229 completed the entire on-boarding process. There were no restrictions on gender, ethnicity, or the participant's location.

Measure

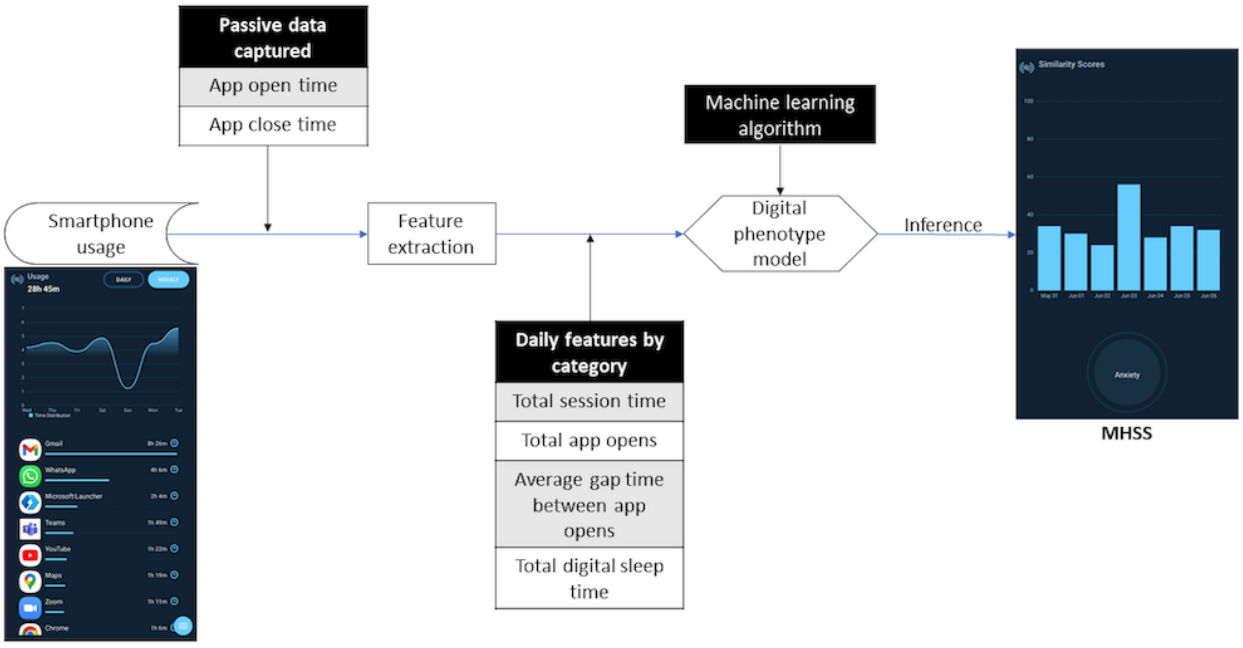
Generalized Anxiety Disorder Screening

The GAD-7 scale [40] is a self-report scale with 7 items for screening nonspecific anxiety in primary care settings. It also indicates the severity of GAD. The items of the scale are rated on a Likert scale ranging from "0=Not at all" to "3=Nearly every day." The scores range from 0 to 21. This questionnaire has good psychometric properties within community and psychiatric samples [41] and has also been established in previous research [42].

Digital Data Collection Through Behavidence

Behavidence [43] is a mental health screening app that passively collects personal smartphone device usage. The app works as a digital profiling solution and can be downloaded from the Google Play Store. There is zero response burden and no collection of any identifiable information. The app was developed for smartphones running Android version 5 or higher. It requires internet connectivity to receive outcomes of data analysis but does not require an active internet connection to collect the data. As the app runs in the background, the participant must provide "Battery Optimization" and "Usage Data Access" permission, obtained during the log-in process. The main screen of the app displays a Mental Health Similarity Score (MHSS), which is inferred from the user's digital behavior. The MHSS displays how similar the user's digital behavior is to someone else's digital behavior who has a diagnosis of anxiety. The similarity score is generated once every 24 hours and has a range of 0%-100%. The app also shows the user their weekly history of daily similarity scores. The workflow of the solution is shown in Figure 1. Data access is managed by multifactor federated authentication and controlled through role-based privileges. Policies are created to manage access for each user, user group, or role. The data pipeline is encrypted end-to-end and orchestrated under enterprise-grade privacy and compliance certification. Data are protected while in-transit via secure socket layer/transport layer security (SSL/TLS) and client-side encryption. Server-side encryption with managed keys is used before storing the data. The application is Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Rule (GDPR) compliant.

Figure 1. The Behavidence solution workflow demonstrates key steps in the creation of a mental health similarity score for anxiety. MHSS: Mental Health Similarity Score.



Data Mining

App Categorization

The total number of apps used by the participants in this study exceeded 50,000 unique apps. To be able to understand and measure features related to each app, we categorized them into 11 categories as follows: Category 0 for nonofficial or unregulated apps, Category 1 for social interaction apps, Category 2 for passive information consumption apps, Category 3 for active messaging and communications apps, Category 4 for educational apps, Category 5 for navigation utilities, Category 6 for general utilities, Category 7 for recreational and photo processing apps, Category 8 for commerce apps, Category 9 for health and fitness–related apps, Category 10 for games, and lastly, Category 11 for miscellaneous.

Feature Extraction Using Passive Smartphone Data

Passive collection of raw nonidentifiable smartphone data starts after the user completes the GAD-7 questionnaire. Seven days of retrograde data are automatically available after a new user log-in, and data are continuously streamed to the back end until the user logs out or deletes the app. The raw data collected include the time in milliseconds of Coordinated Universal Time (UTC) in which a user opens a particular app and the time a user closes that app. From these raw data, behavioral insights used as features for the machine learning algorithms are drawn on a 24-hour basis. For example, the total session time on a phone is calculated by summing the total number of milliseconds the user spends on each app he/she opens, between 12 AM in the user's local time zone to 11:59 PM that day. Incomplete 24-hour data are omitted from the feature engineering process and may be attributed to network disconnection of the user's Android device. No users in this study had gaps of incomplete 24-hour data within consecutive days of collection. Mobile apps were also binned into specific app categories (see the "App Categorization" section) for further insights into digital behavior. Frequency and duration of each app category are calculated daily to indicate where the user spends the most time on their mobile device (ie, shopping, gaming, online dating, communication). Therefore, a total of 34 features were extracted from the original raw data (full list of features are listed in Multimedia Appendix 1).

Data Preparation and Model Setup

A single independent observation in this study constituted 24 hours (user's time zone) of raw data transmitted by the Behavidence App to the back end secure cloud system. Therefore, an individual with anxiety that had 15 days of full passively acquired data was considered to have 15 separate anxiety-labeled observations. To evaluate the models, we reported on different accuracy metrics using 5-fold cross-validation. With 5-fold cross-validation, the data set was split into 5 groups where models were trained on 4 groups and validated on the left-out group. The process was repeated 5 times so that each sample was used for training and validation only once. The Amazon Web Services platform (Amazon.com, Inc.) was used as data storage while the data processing, feature engineering, model training, and poststatistical analysis were written in Python 3.8 programming language (Python Software Foundation). Packages used include scipy, stats models, net neurotools, and scikit-learn.

Modeling and Postanalysis

Machine Learning to Predict Generalized Anxiety

To explore the efficacy of digital behavioral markers in detecting generalized anxiety, regression and classification models were implemented. First, a random forest algorithm was used to create a nonlinear multiple regression fit for the passive digital data corresponding to the total possible score of 21 for the GAD-7 scale. The purpose of this model was to infer what GAD-7 score a user would obtain based on his/her phone usage. For the classification models, 4 different machine learning algorithms were compared to produce the highest overall prediction accuracy. The algorithms compared include random forest, K-nearest neighbors, logistic regression, and XGBoost. The multiclass GAD-7 model is intended to classify participants who scored 15+ (severe), 10-14 (moderate), 5-9 (mild), and <5 (no diagnosis) to detect the progression into severe anxiety. The binary GAD-7 model is intended to classify participants who scored 15+ (severe) on the GAD-7 against those who scored <5 (ie, having no indication of anxiety).

Correlation-Based Analysis

Further analysis on specific items from the GAD-7 was conducted to determine which symptoms of anxiety can be understood from the passively collected digital data. Each of the 7 questions was tested against the MHSS obtained from the top-performing GAD-7 model and calculated on the day each user answered the questionnaire. This testing was performed to determine the existence of a relationship between the digital behaviors collected from the Behavidence app and each question of GAD-7. Nonparametric permutation tests were performed to determine the significance of the Pearson correlation, with the number of permutations set to 1000. Permutation testing was used to better estimate the population's distribution, by not assuming a normal distribution (nonparametric), and to ultimately determine extremities more accurately, by leveraging resampling, so that *P* values indicate the true probability that the Pearson correlation coefficient calculated is not by chance. As the MHSS is derived from the 34 passive digital features, further correlation between specific items from the GAD-7 questionnaire and each of the features was assessed to determine whether the digital biomarker in this study could be mapped to the symptoms of GAD that the specific items are targeting.

Ethics Approval

The advertisement, informed consent, and the study protocol were approved by the independent Western Institutional Board Copernicus Group (WIRB-CG) institutional review board (Approval Number 20216225).

Results

Participants

Self-reported demographic data from the 229 participants (Table 1) show that 85 (37.1%) identified as females, 142 (62%) identified as males, and 2 (0.9%) identified as nonbinary or preferred not to disclose their gender. For the participants' age

distribution, 102 (44.5%) were aged between 18 and 25, 66 (28.8%) between 26 and 35, 56 (24.5%) between 36 and 55, and 5 (2.2%) between 56 and 64. A majority of the participants that completed the questionnaire were of Asian race (104/229, 45.4%), and had education levels between some college diploma and a bachelor's degree (158/229, 69%). The participants in this study were from different locations around the globe. Most were in Asia (84/229, 36.7%) followed by Africa (76/229, 33.2%). The remaining participants were from America, Europe, and Australia.

Table 1. Demographic data of the participants who answered the GAD-7[a] questionnaire (n=229).

Category	Values, n (%)
Age, years	
18-25	102 (44.5)
26-35	66 (28.8)
36-55	56 (24.5)
56-64	5 (2.2)
Gender	
Male	142 (62.0)
Female	85 (37.1)
Prefer not to say	2 (0.9)
Race	
Asian	104 (45.4)
Black (African/Caribbean)	40 (17.5)
White	61 (26.6)
Mixed	11 (4.8)
Other/prefer not to say	13 (5.7)
Education	
Lower secondary/middle school (grades 7-9)	2 (0.9)
Higher secondary (grades 10-12)	35 (15.3)
Some college/university/diploma	74 (32.3)
Bachelor's degree	84 (36.7)
Master's degree	28 (12.2)
Professional/PhD	6 (2.6)
Time zone	
Africa	76 (33.2)
Americas	9 (3.9)
Asia	84 (36.7)
Australia	1 (0.4)
Europe	13 (5.7)
Other[b]	46 (20.1)

[a]GAD-7: 7-item Generalized Anxiety Disorder Scale.

[b]All the other time zones that were unspecified.

GAD-7 Distribution Among Participants

Table 2 represents the distribution of the 229 recruits and their GAD-7 scoring. The GAD-7 was completed at the start of recruitment at a single time point during this study, which spanned from October 2021 to January 2022. The distribution of the GAD-7 scores was as follows: 23/229 (10%) were none with GAD-7 scoring less than 5, while 206/229 (89.9%) showed signs of anxiety by scoring between "mild" and "severe." The mean GAD-7 score was 11.8 (SD 5.7).

Table 2. Distribution of participants' contribution to the GAD-7[a] responses (n=229).

GAD-7 category and scores	Participants, n (%)
None	23 (10)
0	10
1	2
2	3
3	3
4	5
Mild	61 (26.6)
5	13
6	6
7	19
8	11
9	12
Moderate	64 (27.9)
10	10
11	10
12	13
13	17
14	14
Severe	81 (35.4)
15	14
16	10
17	11
18	14
19	7
20	10
21	15

[a]GAD-7: 7-item Generalized Anxiety Disorder Scale.

As seen in Table 3 16% (14/88) of self-reported healthy "none" group participants scored "none" on the GAD-7, whereas the greatest percentage (29/88, 33%) of participants in this group scored "moderate" anxiety. Table 3 also shows that 52% (13/25) of participants with self-reported anxiety had severe anxiety on the GAD-7. Further, 61% (31/51) of participants with self-reported depression had "severe" anxiety and only 2% (1/51) had no signs of anxiety.

Table 3. Distribution of GAD-7[a] scoring categories for self-reported participants.

Self-reported diagnosis	None, n (%)	Mild, n (%)	Moderate, n (%)	Severe, n (%)
a. None (n=88)	14 (16)	15/88 (17)	29/88 (33)	22/88 (25)
b. Anxiety (n=25)	N/A[b]	5/25 (20)	7/25 (28)	13/25 (52)
c. Depression (n=51)	1/51 (2)	6/51 (12)	13/51 (25)	31/51 (61)

[a]GAD-7: 7-item Generalized Anxiety Disorder Scale.

[b]N/A: no participants with a self-reported diagnosis of anxiety scored "none" on the GAD-7 questionnaire.

Evaluation of Models

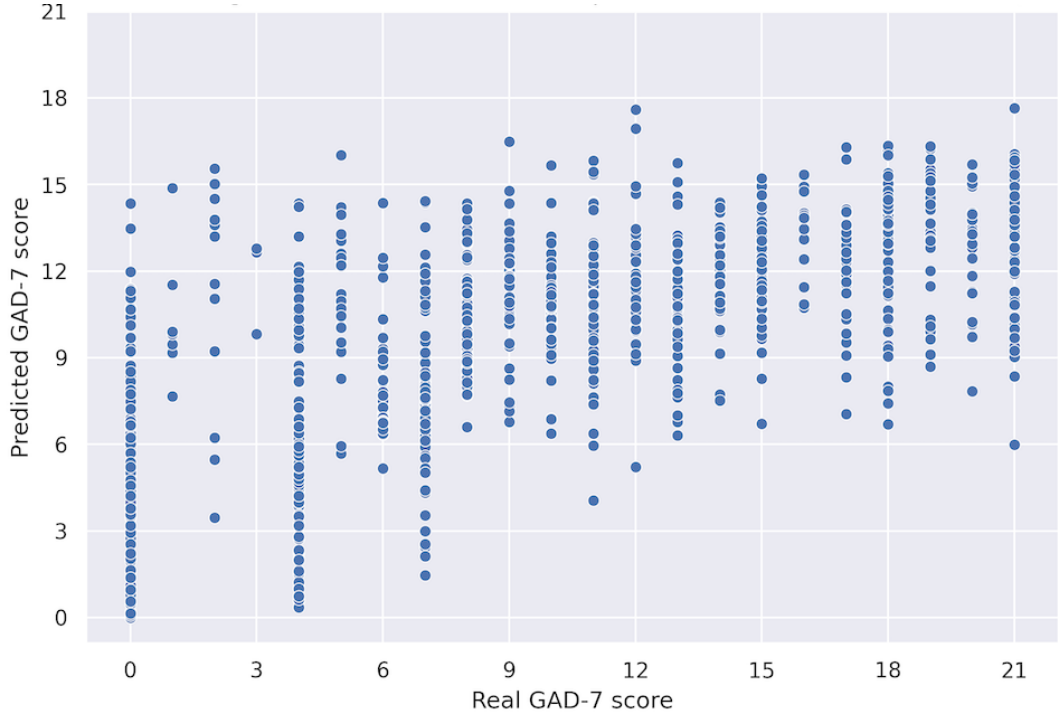
Overview

The aim of the study was to evaluate the accuracy of the MHSS metric to identify GAD. The binary classification XGBoost model achieved a prediction accuracy of 76% compared with 50% by the multiclass classification XGBoost model and regression (root-mean-squared error [RMSE] 4.508). The recall scores for the binary model were 68% for the "none" group and 87% for the "anxiety group." Using the multiclass XGBoost model the best recall scores achieved were 41%, 63%, 38%, and 52% for the "none," "mild," "moderate," and "severe" groups, respectively. The reported results are from the 5-fold cross-validation of data.

Regression Model Assessment

Figure 2 shows the random forest regression model–predicted GAD-7 score plotted against the actual GAD score. The range of predicted values in the lower scores (0-7) is quite high, distributing around 75% of all possible scores. The RMSE for this model is 4.508 with an R^2 value of 0.4282.

Figure 2. Random forest regression: real GAD-7 score versus predicted GAD-7 score (correlation: 0.65597). GAD-7: 7-item Generalized Anxiety Disorder Scale.



Multiclass Classification

The multiclass classification model, trained on all severity group classes, none (GAD-7<5), mild (5≤GAD-7<10), moderate (10≤GAD-7<15), and severe (GAD-7≥20), that achieved the highest prediction accuracy was using XGBoost followed by the random forest algorithm. Result metrics from the 4 algorithm comparisons are presented in Table 4. The GAD-7 multiclass XGBoost model achieved a precision of 40%-62%, recall of 38%-63%, F_1-score of 39%-61%, and overall accuracy of 50%. Sensitivity for severe anxiety was 52% and specificity was 74%.

The Gini impurity plot of each feature shows the top features that the multiclass XGBoost model considers when differentiating between all the possible groups (Figure 3). The 3 most important features in this classifier were the number of times "passive information consumption" apps were opened within the 24-hour period (app2_opens), mean session time within a 24-hour period in "passive information consumption" apps (app2), and the number of times "games" apps were opened with session lengths greater than 1 SD from the mean (app10_upper).

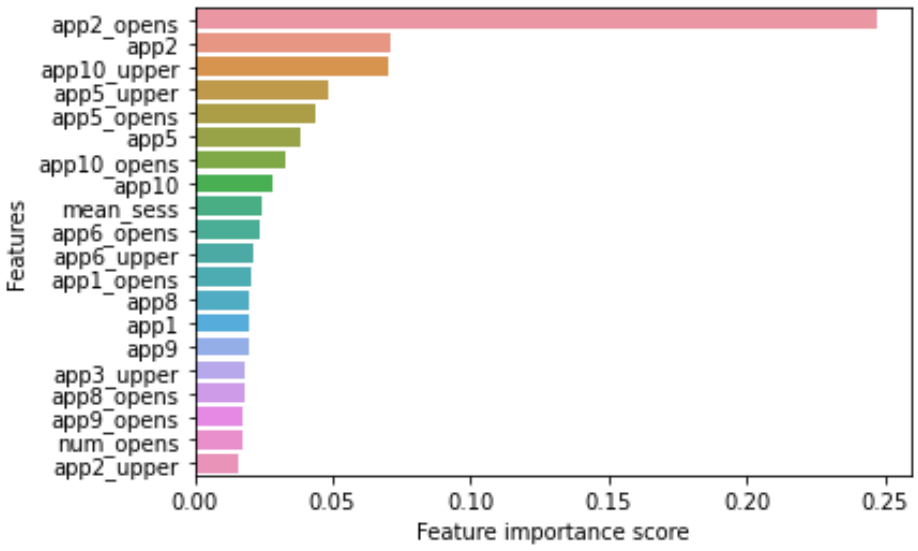
Analysis of variance was performed to determine the difference among means of the 4 different cohorts (ie, none, mild, moderate, and severe) for the top 3 Gini important features. For the feature summing the total number of times "passive information consumption" apps were opened, $F_{4,2619}$=63.40 and P=.44. For the average session time on passive information consumption apps, $F_{4,2619}$=5.23 and P=.002. Finally, for the number of times "games" apps were opened with session lengths greater than 1 SD from the mean, $F_{4,2619}$=60.22 and P=.26. In addition, Tukey post hoc test for pairwise comparison was performed with Cohen d effect size. Detailed results can be found in Multimedia Appendix 2.

Table 4. Multiclass classification accuracy metrics of all algorithms tested in this study (random forest, k-nearest neighbors, logistic regression, XGBoost) using 5-fold cross-validation.

Class	GAD-7[a] multiclass RF model, %	GAD-7 multiclass K-nearest neighbors model, %	GAD-7 multiclass logistic regression model, %	GAD-7 multiclass XGB model, %
Accuracy	48	29	38	50
Area under the curve	69	53	56	71
Precision				
None	64	27	19	62
Mild	58	33	41	60
Moderate	37	27	39	41
Severe	39	27	33	40
Recall				
None	41	22	77	41
Mild	58	34	24	63
Moderate	41	28	0.4	38
Severe	48	29	29	52
F_1-score				
None	50	24	53	50
Mild	58	34	29	61
Moderate	39	28	0.6	39
Severe	43	28	31	45

[a]GAD-7: 7-item Generalized Anxiety Disorder Scale.

Figure 3. Feature importance of the GAD-7 multiclass XGBoost model. GAD-7: 7-item Generalized Anxiety Disorder Scale.



Binary Classification

The random forest classification model, which trained on 2 classes (none vs severe anxiety) and 34 features with the number of trees set to 50, achieved a precision of 79%-70%, recall of 59%-86%, F_1-score of 68%-78%, an overall accuracy of 74%, and area under the curve (AUC) of 78% (Table 5). The binary logistic regression model achieved a precision of 55%-56%, recall of 28%-80%, F_1-score of 37%-66%, an overall accuracy of 55%, and AUC of 57%. The binary K-nearest neighbors model, with k set to 17 according to optimized parametric tuning, achieved a precision of 59%-60%, recall of 46%-73%, F_1-score of 52%-66%, an overall accuracy of 60%, and AUC of 62%. Finally, the binary XGBoost model was the one with the highest accuracy, which achieved a precision of 81%-73%, recall of 68%-87%, F_1-score of 71%-79%, an overall accuracy

of 74%, and AUC of 78%. This model can successfully differentiate between "none" and "severe" anxiety.

In this experiment, the best performing classification algorithm is the XGBoost, which consists of 50 trees that use the Gini criterion to measure the quality of a split with no maximum depth and a minimum of 2 samples per split. The model was further analyzed by plotting Gini impurity values of each feature because this method was used as the splitting criterion of the classification trees when determining the none and severe anxiety groups. As seen in Figure 4, the top 3 passive digital features were mean session time within a 24-hour period in the "passive information consumption" apps (app category 2), mean session time within a 24-hour period in the "health and fitness" apps, and the number of times "passive information consumption" apps were opened within the 24-hour period (app2_opens). The *t* test (unpaired) results indicated statistical significance on all 3 of the top features (Table 6). The effect size ranges from low to high, with the total number of times social interaction apps opened having the greatest effect size (Table 6).

Table 5. Accuracy metrics of all binary classification models trained in this study (random forest, k-nearest neighbors, logistic regression, and XGBoost) using 5-fold cross-validation.

Class	GAD-7[a] binary RF model, %	GAD-7 binary K-nearest neighbors model, %	GAD-7 binary logistic regression model, %	GAD-7 binary XGB model, %
Accuracy	74	60	55	76
AUC[b]	78	62	57	80
Precision				
None	79	59	55	81
Anxiety	70	60	56	73
Recall				
None	59	46	28	68
Anxiety	86	73	80	87
F_1-score				
None	68	52	37	71
Anxiety	78	66	66	79

[a]GAD-7: 7-item Generalized Anxiety Disorder Scale.

[b]AUC: area under the curve.

Figure 4. Feature importance of the GAD-7 Binary XGBoost model. GAD-7: 7-item Generalized Anxiety Disorder scale.

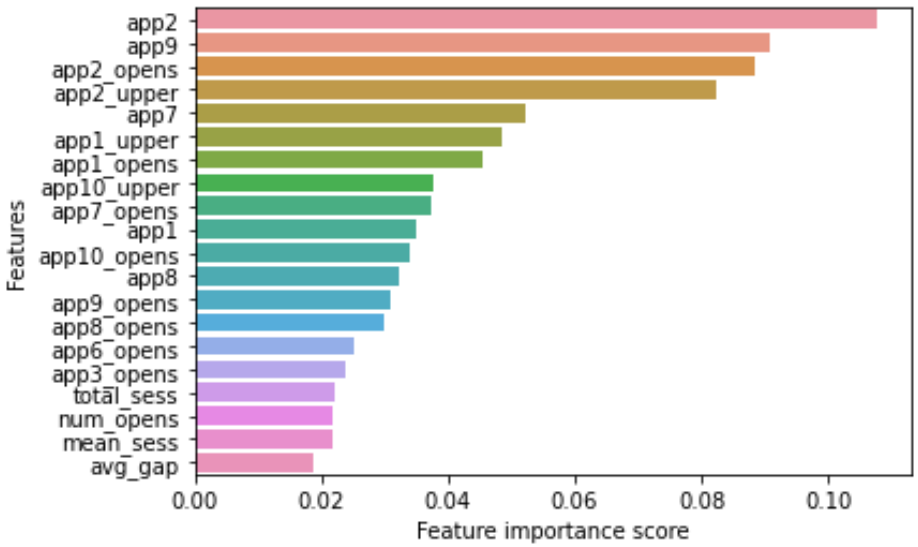


Table 6. Nonparametric *t* tests on the top 3 Gini importance features of the GAD-7[a] binary XGBoost model.

Feature description	None, mean (SD)	Severe, mean (SD)	*P* value	Cohen *d*
App category 2, average session time on passive information consumption apps (minutes)	0.22 (1.03)	0.08 (0.44)	.002	0.18
App category 9, average session time on Health and Fitness apps (minutes)	0.39 (0.93)	0.60 (1.49)	.003	–0.16
App category 2 opens, total number of times passive information consumption apps were opened (count)	0.51 (1.05)	0.43 (2.46)	.45	0.041

[a]GAD-7: 7-item Generalized Anxiety Disorder Scale.

Correlations of GAD-7 Items

Each GAD-7 item was tested using nonparametric permutation testing with Pearson correlation against MHSS on the day that the GAD-7 was filled (Table 7). The highest correlated items belonged to Items 1, 3, and 7: (1) "Feeling nervous, anxious, or on edge" had a correlation of 0.54 (P<.001), (3) "Worrying too much about different things" had a correlation of 0.59 (P<.001), and (7) "Feeling afraid, as if something awful might happen" had a correlation of 0.55 (P<.001).

Table 7. Nonparametric permutation testing with Pearson correlation of GAD-7[a] items against MHSS[b] on the day the questionnaire was filled.

Item	Pearson correlation, *r*	*P* value
1: "Feeling nervous, anxious, or on edge"	0.54	<.001
2: "Not being able to stop or control worrying"	0.5	<.001
3: "Worrying too much about different things"	0.59	<.001
4: "Trouble relaxing"	0.48	<.001
5: "Being so restless that it's hard to sit still"	0.32	<.001
6: "Becoming easily annoyed or irritable"	0.5	<.001
7: "Feeling afraid, as if something awful might happen"	0.55	<.001

[a]GAD-7: 7-item Generalized Anxiety Disorder Scale.

[b]MHSS: Mental Health Similarity Score.

Discussion

Principal Findings

Smartphone technology has certainly become a primary platform not only for communication but also to receive, manage, and share multiple kinds of data. Recently, the application of smartphones and their sensing capabilities have demonstrated huge potential in health information acquisition and analysis [25-30,34-38]. Mining smartphone data to represent digital behavior can be used for delivering informed clinical decisions and early risk stratification of mental health disorders. Through this study, we demonstrate the application of digital phenotyping in the identification and remote monitoring of GAD.

A novel mental behavioral profiling metric called MHSS was derived by engineering 34 digital features to serve as a marker for GAD. This was accomplished using smartphone usage data mined in a passive manner without the use of any private information. The smartphone usage data comprised active app usage time and frequency collected through the Behavidence app for an average period of 14 days per user. A single observation that consists of 24 hours of smartphone usage data had a typical size of 30 KB. During the course of the study, the engagement with the Behavidence app (number of times the app was opened per day) had an average of 0.78%, highlighting the benefit of zero respondent burden. Answering the GAD-7 questionnaire was only for the purpose of training the models and testing its performance. Models created in the study explored the ability of the MHSS to predict the GAD-7 outcome at 3 levels of granularity. The regression model explored the conformance of MHSS to GAD-7 on an individual score level (0-21) and achieved an RMSE of 4.508. The multiclass classification model encoded 4 levels of anxiety severity with an overall accuracy of 50%, whereas the binary classification model distinguished individuals with severe anxiety from the ones without any anxiety with an overall accuracy of 76%.

Although there can be a substantial within-subject variability in scoring across time as mentioned by Meyerhoff et al [28], the reported SD for GAD-7 (3.50) is less than the RMSE achieved in this study. In a clinical use case, the GAD-7 score–based anxiety category is more relevant than the individual scores. Interrater reliability of anxiety disorder diagnosis is shown to have a κ value of 0.20 [44]. A key performance indicator for MHSS would be its ability to differentiate individuals across the anxiety categories with an accuracy over 70%. Each anxiety category (ie, none, mild, moderate, and severe) has a range of 4 points in the GAD-7 scale. As the RMSE in this regression model exceeds this range, this model would result in very low accuracy of anxiety category prediction.

The GAD-7 multiclass model achieved an overall accuracy of 50%, with a sensitivity of 63%, 37%, 41%, and 52% and specificity of 80%, 84%, 93%, and 74% for the none, mild, moderate, and severe classes, respectively. Prior studies

performed in primary care clinics have noted that a cut-off score of 10 or higher on the GAD-7 scale has a sensitivity of 89% and specificity of 82% [45]. Although GAD-7 may be particularly useful in assessing symptom severity, a score of 10 or greater on the GAD-7 is most reliable for identifying cases of GAD. This supports the case for developing a binary classification model as an effective screening tool. With the available number of participants in the study, the statistical power for differentiating participants with severe anxiety from ones without anxiety using the digital phenotype as a marker was the strongest (76%). Based on testing various modeling algorithms including random forest, logistic regression, K-nearest neighbors, and RF, the GAD-7 binary XGBoost model achieved 76% accuracy with a sensitivity of 62% and specificity of 86%. These accuracy levels are higher than published results that use intrusive markers to predict generalized and social anxiety disorder [25], or that have used physiological markers to predict anxiety severity [46]. Along with the accuracy levels, sensitivity and specificity results for the GAD-7 binary model are also higher than studies done by Nemesure et al [47] and Fukazawa et al [48], which used binary classification for prediction of anxiety.

One of the key findings was the higher use of certain app categories such as "passive information consumption apps," "games," and "health and fitness" among participants with anxiety as compared with those without. Feature importance analysis has been performed by various previous studies, and they have demonstrated the usefulness of knowing these predictors [49]. Previous studies have stated various features such as daily screen time [25] as useful predictors. This study highlights certain app categories as important predicting features, allowing a deep dive into the digital usage patterns of people with and without anxiety. Whether the increased usage of such apps is a result or a cause of elevated anxiety is a topic for further exploration.

The correlation analysis performed between the items of the GAD-7 scale found that the highest correlated items were 1, 3, and 7. This has been a very interesting finding because the 2-factor structure of the GAD-7 scale has been suggested in previous studies such as Beard and Björgvinsson [50], where Items 1, 2, 3, and 7 belonged to the cognitive and emotional component of anxiety and 4, 5, and 6 to the somatic component. This points to the result that machine learning algorithms employed to generate MHSS are more sensitive in picking up the emotional/cognitive component of anxiety.

Study Implications

The MHSS for anxiety has the potential to serve as a complementary continuous metric to the GAD-7 questionnaire as well as clinical assessment of anxiety disorder. This metric has the advantage of being able to monitor daily anxiety levels with no respondent burden. This enables the use of smartphone-based sensing to overcome any "state-of-mind" biases. Given the metric's sensitivity to the emotional/cognitive component of anxiety, it can help in overcoming those undiagnosed cases where somatic symptoms of anxiety result in a conflict in diagnosis. This is especially useful in cases where there is an overlap of physical symptoms (shortness of breath or palpitations) and cognitive symptoms (such as insomnia, restlessness) as well as an overlap with depression [9,19]. Another potential use for MHSS is outlining and differentiating the risk of comorbidities. Anxiety disorders are mostly comorbid with depression. A recent study using the same Behavidence research app was able to predict depression severity with the MHSS for depression. Choudhary et al [26] found that machine learning models that generated an MHSS for depression had high accuracy metrics (≥89%) and were able to distinguish between users with depression and those without. Coupled with the findings of this study, MHSS can distinguish between comorbid depression and anxiety, thereby improving clinical decision making.

Limitations and Future Work of the Study

One of the limitations of the study was that the GAD-7 questionnaire was collected at only 1 time point during the study. In this study the sample size was average, with unequal amounts of gender proportions and education background, which can affect the generalizability of the study, as GAD is a very commonly observed phenomenon. Although the study had almost equal proportions of mild, moderate, and severe groups of anxiety, this was an online recruited sample. With accurate model metrics, further studies should aim for having clinical samples and populations. Therefore, future models should focus on recruiting larger sample sizes and clinical populations to further test the applicability of such findings. Although the machine learning models indicate a higher accuracy of the GAD-7 binary model, the MHSS may have different thresholds for various levels of anxiety severity, which should be subjected to further research. Given the existence of comorbidities, particularly depression, a dedicated study to assess the correlation between MHSS for depression and MHSS for anxiety could generate valuable insights and shed light on how different interventions may be impactful.

Conclusion

The lack of access to mental health care can be addressed through the ubiquitously available smartphone and the development of passive and widely available screening technologies for detecting the most common mental health disorders. Objective smartphone-collected data contain enough information about an individual's digital behavior to infer his/her mental states and screen for anxiety, and is a technology that provides remote, longitudinal, and continuous monitoring as an integrative and agile solution. Machine learning serves as an effective technique to mine such big data to derive accurate biomarkers for mental health conditions such as anxiety.

Acknowledgments

The study was funded by Behavidence Inc.

Authors' Contributions

NT and SC wrote the paper. NT, SA, and UN performed data analysis for the study. JE, GS, and RC edited the paper and offered their expertise.

Conflicts of Interest

All the authors have jointly developed the Behavidence Research app and are now employed at Behavidence Inc.

Multimedia Appendix 1
The list of 34 features mined from the data.
[PDF File (Adobe PDF File), 48 KB - medinform_v10i8e38943_app1.pdf]

Multimedia Appendix 2
Tukey post hoc test for pairwise comparison of the top 3 digital features from the GAD-7 multiclass XGBoost model with Cohen *d* effect size. GAD-7: 7-item Generalized Anxiety Disorder scale.
[DOCX File , 14 KB - medinform_v10i8e38943_app2.docx]

References

1. Friedrich M. Depression Is the Leading Cause of Disability Around the World. JAMA 2017 Apr 18;317(15):1517. [doi: 10.1001/jama.2017.3826] [Medline: 28418490]
2. Craske MG, Stein MB, Eley TC, Milad MR, Holmes A, Rapee RM, et al. Anxiety disorders. Nat Rev Dis Primers 2017 May 04;3(1):17024. [doi: 10.1038/nrdp.2017.24]
3. Weisberg RB, Beard C, Pagano ME, Maki KM, Culpepper L, Keller MB. Impairment and Functioning in a Sample of Primary Care Patients With Generalized Anxiety Disorder. Prim. Care Companion J. Clin. Psychiatry 2010 Oct 07:PCC.09m00890. [doi: 10.4088/pcc.09m00890blu]
4. Wilmer MT, Anderson K, Reynolds M. Correlates of Quality of Life in Anxiety Disorders: Review of Recent Research. Curr Psychiatry Rep 2021 Oct 06;23(11):77 [FREE Full text] [doi: 10.1007/s11920-021-01290-4] [Medline: 34613508]
5. Bandelow B, Michaelis S. Epidemiology of anxiety disorders in the 21st century. Dialogues in Clinical Neuroscience 2022 Apr 01;17(3):327-335. [doi: 10.31887/dcns.2015.17.3/bbandelow]
6. Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):617-627 [FREE Full text] [doi: 10.1001/archpsyc.62.6.617] [Medline: 15939839]
7. Roberge P, Normand-Lauzière F, Raymond I, Luc M, Tanguay-Bernard M, Duhoux A, et al. Generalized anxiety disorder in primary care: mental health services use and treatment adequacy. BMC Fam Pract 2015 Oct 22;16(1):146 [FREE Full text] [doi: 10.1186/s12875-015-0358-y] [Medline: 26492867]
8. Castro-Rodríguez JI, Olariu E, Garnier-Lacueva C, Martín-López LM, Pérez-Solà V, Alonso J, INSAyD investigators. Diagnostic accuracy and adequacy of treatment of depressive and anxiety disorders: A comparison of primary care and specialized care patients. J Affect Disord 2015 Feb 01;172:462-471. [doi: 10.1016/j.jad.2014.10.020] [Medline: 25451451]
9. Vermani M, Marcus M, Katzman MA. Rates of Detection of Mood and Anxiety Disorders in Primary Care. Prim. Care Companion CNS Disord 2011 Apr 28:PCC.10m01013. [doi: 10.4088/pcc.10m01013]
10. Fletcher T, Helm A, Vaghani V, Kunik M, Stanley M, Singh H. Identifying psychiatric diagnostic errors with the Safer Dx Instrument. Int J Qual Health Care 2020 Jul 20;32(6):405-411. [doi: 10.1093/intqhc/mzaa066] [Medline: 32671387]
11. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Arch Intern Med 2010 Jun 28;170(12):1015-1021. [doi: 10.1001/archinternmed.2010.146] [Medline: 20585065]
12. Fletcher T, Hundt N, Kunik M, Singh H, Stanley M. Accuracy of Anxiety Disorder Not Otherwise Specified Diagnosis in Older Veterans. J Psychiatr Pract 2019 Sep;25(5):358-364. [doi: 10.1097/PRA.0000000000000408] [Medline: 31505520]
13. Baldwin DS, Allgulander C, Bandelow B, Ferre F, Pallanti S. An international survey of reported prescribing practice in the treatment of patients with generalised anxiety disorder. World J Biol Psychiatry 2012 Oct 07;13(7):510-516. [doi: 10.3109/15622975.2011.624548] [Medline: 22059936]
14. Fernández A, Haro JM, Codony M, Vilagut G, Martínez-Alonso M, Autonell J, et al. Treatment adequacy of anxiety and depressive disorders: primary versus specialised care in Spain. J Affect Disord 2006 Nov;96(1-2):9-20. [doi: 10.1016/j.jad.2006.05.005] [Medline: 16793140]
15. Phillips J. Detecting diagnostic error in psychiatry. Diagnosis (Berl) 2014 Jan 01;1(1):75-78 [FREE Full text] [doi: 10.1515/dx-2013-0032] [Medline: 29539971]
16. Coryell W. Diagnostic instability: how much is too much? Am J Psychiatry 2011 Nov;168(11):1136-1138. [doi: 10.1176/appi.ajp.2011.11081191] [Medline: 22193597]
17. Johnson EM, Coles ME. Failure and delay in treatment-seeking across anxiety disorders. Community Ment Health J 2013 Dec 29;49(6):668-674. [doi: 10.1007/s10597-012-9543-9] [Medline: 23054147]

18. Wu Z, Fang Y. Comorbidity of depressive and anxiety disorders: challenges in diagnosis and assessment. Shanghai Arch Psychiatry 2014 Aug;26(4):227-231 [FREE Full text] [doi: 10.3969/j.issn.1002-0829.2014.04.006] [Medline: 25317009]
19. Kartal M. In: Ãgnes S, editor. Challenges and Opportunities in Diagnosis and Management of Generalized Anxiety Disorder in Primary Care, Anxiety and Related Disorders. Rijeka, Croatia: InTech; Aug 2011.
20. Lewis G. Observer bias in the assessment of anxiety and depression. Soc Psychiatry Psychiatr Epidemiol 1991;26(6):265-272. [doi: 10.1007/bf00789218]
21. Nordahl H, Plummer A, Wells A. Predictors of Biased Self-perception in Individuals with High Social Anxiety: The Effect of Self-consciousness in the Private and Public Self Domains. Front Psychol 2017 Jul 04;8:1126 [FREE Full text] [doi: 10.3389/fpsyg.2017.01126] [Medline: 28725207]
22. Newson J, Pastukh V, Thiagarajan T. Poor Separation of Clinical Symptom Profiles by DSM-5 Disorder Criteria. Front Psychiatry 2021;12:775762 [FREE Full text] [doi: 10.3389/fpsyt.2021.775762] [Medline: 34916976]
23. Zimmerman M, Ellison W, Young D, Chelminski I, Dalrymple K. How many different ways do patients meet the diagnostic criteria for major depressive disorder? Compr Psychiatry 2015 Jan;56:29-34. [doi: 10.1016/j.comppsych.2014.09.007] [Medline: 25266848]
24. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019 Sep 6;2(1):88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
25. Di Matteo D, Fotinos K, Lokuge S, Mason G, Sternat T, Katzman MA, et al. Automated Screening for Social Anxiety, Generalized Anxiety, and Depression From Objective Smartphone-Collected Data: Cross-sectional Study. J Med Internet Res 2021 Aug 13;23(8):e28918 [FREE Full text] [doi: 10.2196/28918] [Medline: 34397386]
26. Choudhary S, Thomas N, Ellenberger J, Srinivasan G, Cohen R. A Machine Learning Approach for Detecting Digital Behavioral Patterns of Depression Using Nonintrusive Smartphone Data (Complementary Path to Patient Health Questionnaire-9 Assessment): Prospective Observational Study. JMIR Form Res 2022 May 16;6(5):e37736 [FREE Full text] [doi: 10.2196/37736] [Medline: 35420993]
27. Yang M, Tang J, Wu Y, Liu Z, Hu X, Hu B. A Behaviour Patterns Extraction Method for Recognizing Generalized Anxiety Disorder. New York, NY: IEEE; 2020 Mar 1 Presented at: 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM); March 1-2, 2021; Shenzhen, China URL: https://sci-hub.se/10.1109/HEALTHCOM49281.2021.9398995 [doi: 10.1109/healthcom49281.2021.9398995]
28. Meyerhoff J, Liu T, Kording KP, Ungar LH, Kaiser SM, Karr CJ, et al. Evaluation of Changes in Depression, Anxiety, and Social Anxiety Using Smartphone Sensor Features: Longitudinal Cohort Study. J Med Internet Res 2021 Sep 03;23(9):e22844 [FREE Full text] [doi: 10.2196/22844] [Medline: 34477562]
29. Jacobson NC, Summers B, Wilhelm S. Digital Biomarkers of Social Anxiety Severity: Digital Phenotyping Using Passive Smartphone Sensors. J Med Internet Res 2020 May 29;22(5):e16875 [FREE Full text] [doi: 10.2196/16875] [Medline: 32348284]
30. MacLeod L, Suruliraj B, Gall D, Bessenyei K, Hamm S, Romkey I, et al. A Mobile Sensing App to Monitor Youth Mental Health: Observational Pilot Study. JMIR Mhealth Uhealth 2021 Oct 26;9(10):e20638 [FREE Full text] [doi: 10.2196/20638] [Medline: 34698650]
31. Sheikh M, Qassem M, Kyriacou P. Wearable, Environmental, and Smartphone-Based Passive Sensing for Mental Health Monitoring. Front Digit Health 2021;3:662811 [FREE Full text] [doi: 10.3389/fdgth.2021.662811] [Medline: 34713137]
32. Boukhechba M, Chow P, Fua K, Teachman BA, Barnes LE. Predicting Social Anxiety From Global Positioning System Traces of College Students: Feasibility Study. JMIR Ment Health 2018 Jul 04;5(3):e10101 [FREE Full text] [doi: 10.2196/10101] [Medline: 29973337]
33. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: UbiComp '14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. New York, NY: Association for Computing Machinery; 2014 Sep Presented at: UbiComp '14: The 2014 ACM Conference on Ubiquitous Computing; September 13-17, 2014; Seattle, WA p. 3-14. [doi: 10.1145/2632048.2632054]
34. Torous J, Chan SR, Yee-Marie Tan S, Behrens J, Mathew I, Conrad EJ, et al. Patient Smartphone Ownership and Interest in Mobile Apps to Monitor Symptoms of Mental Health Conditions: A Survey in Four Geographically Distinct Psychiatric Clinics. JMIR Ment Health 2014;1(1):e5 [FREE Full text] [doi: 10.2196/mental.4004] [Medline: 26543905]
35. Mendes JPM, Moura IR, Van de Ven P, Viana D, Silva FJS, Coutinho LR, et al. Sensing Apps and Public Data Sets for Digital Phenotyping of Mental Health: Systematic Review. J Med Internet Res 2022 Feb 17;24(2):e28735 [FREE Full text] [doi: 10.2196/28735] [Medline: 35175202]
36. Gao Y, Li A, Zhu T, Liu X, Liu X. How smartphone usage correlates with social anxiety and loneliness. PeerJ 2016;4:e2197 [FREE Full text] [doi: 10.7717/peerj.2197] [Medline: 27478700]
37. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health 2020 Nov 30;23(4):161-166. [doi: 10.1136/ebmental-2020-300180] [Medline: 32998937]

38. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. J Med Internet Res 2015 Jul 15;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
39. Gelinas L, Pierce R, Winkler S, Cohen IG, Lynch HF, Bierer BE. Using Social Media as a Research Recruitment Tool: Ethical Issues and Recommendations. Am J Bioeth 2017 Mar;17(3):3-14 [FREE Full text] [doi: 10.1080/15265161.2016.1276644] [Medline: 28207365]
40. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
41. Rutter LA, Brown TA. Psychometric Properties of the Generalized Anxiety Disorder Scale-7 (GAD-7) in Outpatients with Anxiety and Mood Disorders. J Psychopathol Behav Assess 2017 Mar;39(1):140-146 [FREE Full text] [doi: 10.1007/s10862-016-9571-9] [Medline: 28260835]
42. Johnson S, Ulvenes P, Øktedalen T, Hoffart A. Psychometric Properties of the General Anxiety Disorder 7-Item (GAD-7) Scale in a Heterogeneous Psychiatric Sample. Front Psychol 2019;10:1713 [FREE Full text] [doi: 10.3389/fpsyg.2019.01713] [Medline: 31447721]
43. Mental Health Application. Behavidence. 2022. URL: https://www.behavidence.com/ [accessed 2022-08-09]
44. Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, Tamminga CA, et al. The initial field trials of DSM-5: new blooms and old thorns. Am J Psychiatry 2013 Jan;170(1):1-5. [doi: 10.1176/appi.ajp.2012.12091189] [Medline: 23288382]
45. Zhong Q, Gelaye B, Zaslavsky AM, Fann JR, Rondon MB, Sánchez SE, et al. Diagnostic Validity of the Generalized Anxiety Disorder - 7 (GAD-7) among Pregnant Women. PLoS One 2015;10(4):e0125096 [FREE Full text] [doi: 10.1371/journal.pone.0125096] [Medline: 25915929]
46. Shaukat-Jali R, van Zalk N, Boyle DE. Detecting Subclinical Social Anxiety Using Physiological Data From a Wrist-Worn Wearable: Small-Scale Feasibility Study. JMIR Form Res 2021 Oct 07;5(10):e32656 [FREE Full text] [doi: 10.2196/32656] [Medline: 34617905]
47. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep 2021 Jan 21;11(1):1980 [FREE Full text] [doi: 10.1038/s41598-021-81368-4] [Medline: 33479383]
48. Fukazawa Y, Ito T, Okimura T, Yamashita Y, Maeda T, Ota J. Predicting anxiety state using smartphone-based passive sensing. J Biomed Inform 2019 May;93:103151 [FREE Full text] [doi: 10.1016/j.jbi.2019.103151] [Medline: 30880254]
49. Richter T, Fishbain B, Markus A, Richter-Levin G, Okon-Singer H. Using machine learning-based analysis for behavioral differentiation between anxiety and depression. Sci Rep 2020 Oct 02;10(1):16381 [FREE Full text] [doi: 10.1038/s41598-020-72289-9] [Medline: 33009424]
50. Beard C, Björgvinsson T. Beyond generalized anxiety disorder: psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. J Anxiety Disord 2014 Aug;28(6):547-552. [doi: 10.1016/j.janxdis.2014.06.002] [Medline: 24983795]

Abbreviations

GAD: generalized anxiety disorder
GAD-7: 7-item Generalized Anxiety Disorder Scale
GDPR: General Data Protection Rule
HIPAA: Health Insurance Portability and Accountability Act
MHSS: Mental Health Similarity Score
RMSE: root-mean-squared error
SSL/TLS: secure socket layer/transport layer security
UTC: Coordinated Universal Time

Edited by C Lovis, J Hefner; submitted 25.04.22; peer-reviewed by J Kim, F Rudzicz, FM Calisto; comments to author 11.05.22; revised version received 11.07.22; accepted 01.08.22; published 30.08.22.

Please cite as:
Choudhary S, Thomas N, Alshamrani S, Srinivasan G, Ellenberger J, Nawaz U, Cohen R
A Machine Learning Approach for Continuous Mining of Nonidentifiable Smartphone Data to Create a Novel Digital Biomarker Detecting Generalized Anxiety Disorder: Prospective Cohort Study
JMIR Med Inform 2022;10(8):e38943
URL: https://medinform.jmir.org/2022/8/e38943
doi:10.2196/38943
PMID:36040777

Original Paper

Effect of Applying a Real-Time Medical Record Input Assistance System With Voice Artificial Intelligence on Triage Task Performance in the Emergency Department: Prospective Interventional Study

Ara Cho[1], MD; In Kyung Min[2], MSc; Seungkyun Hong[3], PhD; Hyun Soo Chung[1], MD, PhD; Hyun Sim Lee[4], PhD; Ji Hoon Kim[1], MD, MPH, PhD

[1]Department of Emergency Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Department of Research Affairs, Biostatistics Collaboration Unit, Yonsei University College, Seoul, Republic of Korea

[3]CONNECT-AI Research Center, Yonsei University College of Medicine, Seoul, Republic of Korea

[4]Department of Emergency Nursing, Yonsei University Health System, Seoul, Republic of Korea

Corresponding Author:
Ji Hoon Kim, MD, MPH, PhD
Department of Emergency Medicine, Yonsei University College of Medicine
50 Yonsei-ro
Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 2465
Email: jichoon81@yuhs.ac

Abstract

Background: Natural language processing has been established as an important tool when using unstructured text data; however, most studies in the medical field have been limited to a retrospective analysis of text entered manually by humans. Little research has focused on applying natural language processing to the conversion of raw voice data generated in the clinical field into text using speech-to-text algorithms.

Objective: In this study, we investigated the promptness and reliability of a real-time medical record input assistance system with voice artificial intelligence (RMIS-AI) and compared it to the manual method for triage tasks in the emergency department.

Methods: From June 4, 2021, to September 12, 2021, RMIS-AI, using a machine learning engine trained with 1717 triage cases over 6 months, was prospectively applied in clinical practice in a triage unit. We analyzed a total of 1063 triage tasks performed by 19 triage nurses who agreed to participate. The primary outcome was the time for participants to perform the triage task.

Results: The median time for participants to perform the triage task was 204 (IQR 155, 277) seconds by RMIS-AI and 231 (IQR 180, 313) seconds using manual method; this difference was statistically significant (P<.001). Most variables required for entry in the triage note showed a higher record completion rate by the manual method, but in the recording of additional chief concerns and past medical history, RMIS-AI showed a higher record completion rate than the manual method. Categorical variables entered by RMIS-AI showed less accuracy compared with continuous variables, such as vital signs.

Conclusions: RMIS-AI improves the promptness in performing triage tasks as compared to using the manual input method. However, to make it a reliable alternative to the conventional method, technical supplementation and additional research should be pursued.

(JMIR Med Inform 2022;10(8):e39892) doi:10.2196/39892

KEYWORDS

voice recognition; artificial intelligence; natural language processing; emergency department; triage

Introduction

An essential role of a hospital emergency department (ED) is to prioritize treatment for patients according to urgency and symptom severity [1]. This role is related to the nature of ED work, where unpredictable situations often occur, and resources are limited owing to crowding [2,3]. Because emergency care demands higher efficacy to manage growing patient volumes, a prompt and evidence-based triage system is required to provide safe and optimal care [4]. Most EDs are equipped with a "triage system" that immediately classifies the severity of a patient's symptoms in the period between patient arrival and start of clinical steps by ED physicians [5]. Initial severity classification includes checking vital signs and recording patient history by conversing with the patient or guardian [6].

Because the results derived through the triage system must be immediately recorded and shared with the medical staff in charge of the next process, a prompt triage system is crucial for an efficient ED. In addition, the results from the triage system are reported to have a significant influence on clinical outcomes [7-10]. Therefore, the accuracy of the triage process is also important for the safe operation of an ED. However, the existing triage system is mostly operated by medical staff rather than physicians, and there may be bias due to the subjective measurement [5]. In addition, because the time required in the triage unit has been prolonged because of the COVID-19 outbreak, rapid and reliable patient classification is threatened in EDs [11].

Recent advances in machine learning and natural language processing (NLP) are a prominent development in health informatics and are relevant in emergency medicine [12]. Although NLP has been established as an important tool when using unstructured text data, most studies in the medical field have been limited to a retrospective analysis of text entered manually by humans on electronic medical records [6,13-18]. There is little research that addresses applying NLP to the conversion of raw voice data generated in the clinical field into text using speech-to-text (STT) algorithms [1,19]. Therefore, the aim of this study was to investigate the promptness and reliability of a real-time record input assistance system developed with STT and NLP technology and compare it to the manual method used by triage medical staff who perform time-critical tasks in the ED.

Methods

Ethics Approval

This study was conducted in accordance with the revised Declaration of Helsinki and was reviewed and approved by the Institutional Review Board of Severance Hospital, South Korea (approval number 4-2020-0598).

Study Setting and Participants

We performed a prospective interventional study. This study was conducted at a Level 1 ED at a tertiary hospital located in northwestern Seoul (the capital city of South Korea), where 90,000 patients visit annually. The hospital's ED is responsible for receiving patients who cannot be stabilized in this catchment area. Participants were recruited through an official announcement period from November 1, 2020, to the end of January 2021. Among the nurses performing triage work in the hospital's ED, 19 nurses who listened to the contents and process of the study voluntarily agreed to participate in the study. They had more than three years of ED work experience. Exclusion criteria included candidates who (1) withdrew their intention to participate, or (2) had physical symptoms that made it difficult for them to wear a voice recognition microphone. Informed consent was obtained from all participants before enrollment.

Machine Learning Framework

Because conversations in the triage unit contained a large amount of information and noise, a device that can select and record these conversations was needed. A machine learning framework created by Selvas AI Inc (Seoul, Republic of Korea) was used in this study. The voice recognition solution provided by Selvas analyzes sound information and converts it into text, commands, and various forms of information. The application of continuous word recognition engines, which recognize unstructured speech, has expanded to different fields; for example, a speech recognition engine in this study has been exclusively developed for the medical field. In our ED, the triage nurses are supposed to record the results of performing a task in a triage note. This triage note consists of the following items: chief concern, past medical history, the presence of allergic diseases, vital signs such as systolic and diastolic blood pressure, heart rate, respiratory rate, body temperature, and oxygen saturation. To train the engine, triage nurses who agreed to participate in this study performed the clinical practice wearing Bluetooth microphones (Aftershokz Aeropex, AS 800, Aftershokz LLC). Voice recording files that passed through the engine were immediately converted into textual data, without prior editing, and stored as log records. Subsequently, the engine repeatedly trained the NLP to fill the items constituting the triage note using the transcribed textual data. The Bluetooth microphone was selected as a component of a noise-resistant system in accordance with the ED environment where various noises exist, and a mobile recording system was built to ensure its mobility. The Bluetooth microphones, voice recognition software, and systems using computers connected to them were installed in the triage unit, and voice data were recorded during the data collection period. For 6 months, 1717 triage cases were collected, and the machine learning engine was trained to recognize the sound using these voice data, convert it into textual data, and perform the subsequent NLP. Consistent with the current triage note format, the system was trained to classify the chief concern of each patient into 1 of 52 categories, and the past medical history was processed into 13 categories through NLP. In the triage note used in this ED, up to 3 chief concerns and the medical history can be entered. The presence of allergic diseases was configured to be treated as a binary input, and variables representing vital signs were treated as continuous variables.

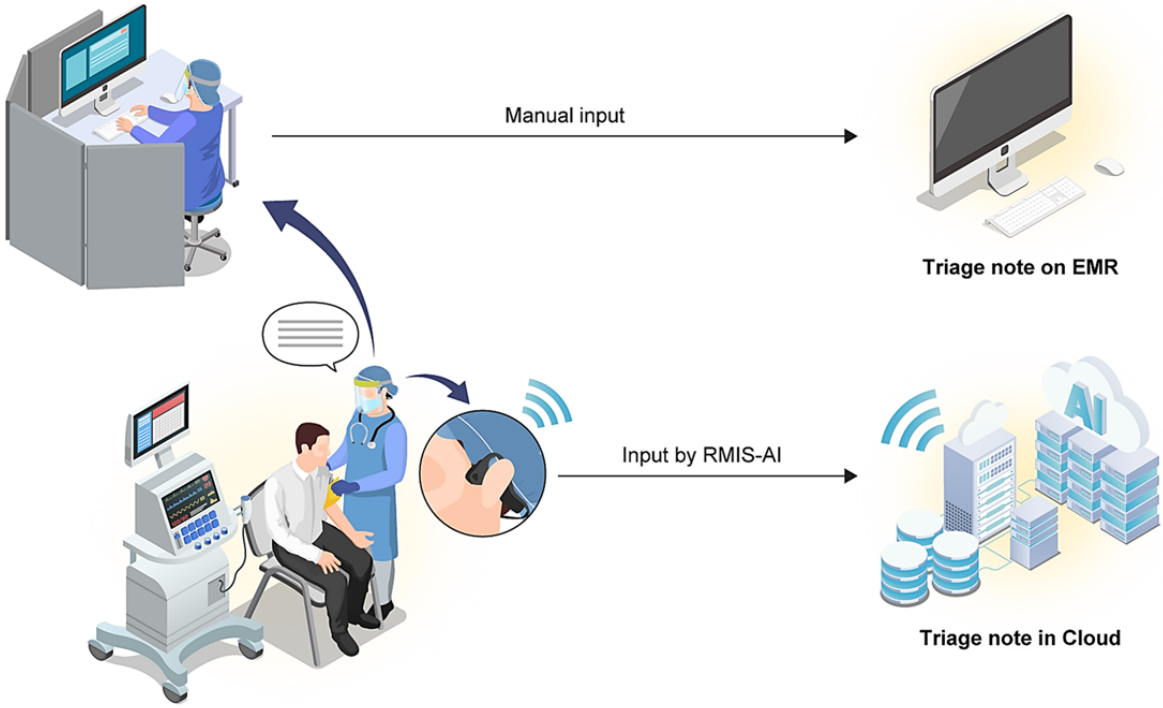
For accurate voice interval detection in a noisy environment, the end-point detection module was optimized in the machine learning engine. By distinguishing various nonstationary noises through continuous adaptive learning for noise coming through the Bluetooth microphones, a deep neural network end-point

detection module was developed with high accuracy in detecting energy-based voice sections of the existing method. The voice interval detection module optimized for the voice environment input to the Bluetooth microphone was advanced, and sound using the collected and processed purified voice database and converted textural data was applied for language model learning.

Study Protocol

From June 4, 2021, to September 12, 2021, a real-time medical record input assistance system with voice artificial intelligence (RMIS-AI) built using a trained engine was prospectively applied to the clinical practice in the triage unit where the patients meet the medical staff for the first time. RMIS-AI is a tool that assists in recording triage notes through voices. In other words, it secures the mobility of a triage nurse by replacing the record input means with voice instead of the desktop computer keyboard. RMIS-AI was implemented on a cloud-based network separate from the hospital electronic medical record (EMR) system. During the study period, participants wearing Bluetooth microphones recorded triage data in the EMR by asking detailed questions to each patient and checked vital signs. Simultaneously, they also recorded the data through RMIS-AI in the same format using their voice. Because the participants used a closed-loop communication method that reconfirmed the meaning of the patient's words and uttered them, the information obtained from the patients could be delivered by the participant's voice rather than the patient's voice. The input process of charting through RMIS-AI was blind to the nurses, and they monitored the EMR input process as usual when performing the triage task. The contents and time of the triage log finally created in both ways were stored in the hospital EMR log and cloud storage, respectively (Figure 1). The data stored in each database were automatically extracted and used for our research.

Figure 1. Two-input process of charting, RMIS-AI (real-time medical record input assistance system with voice artificial intelligence) vs manual input. EMR: electronic medical record.



Outcome Measures

The primary outcome was the time for participants to perform the triage task. It was defined as the time from the patient's arrival at the triage unit to the completion of the triage note. We measured these times using data stored in the hospital EMR for manual input and cloud storage for RMIS-AI. The secondary outcome metrics were the record completion rate and the accuracy of RMIS-AI compared to manual input by EMR.

Statistical Analysis

The sample size was calculated from the mean time taken by performing the triage task in a conventional method for 100 cases before the intervention was started. We considered that the RMIS-AI producing a mean difference of 20 seconds with standard deviation difference of 2 seconds would be considered clinically significant (P<.05, statistical power=0.95). Therefore, the required sample size was calculated to be 952 cases by G-power 3.1.9.7, requiring a total of 1057 triage cases considering a 10% dropout rate. In this paper, categorical variables are presented as counts and percentage. Continuous data are presented as mean or median and SD or interquartile range. The Mann-Whitney U test was used to identify the differences of primary outcome between the 2 groups. Differences in record completion rates between the 2 methods were compared using the McNemar test. The result was considered statistically significant at P<.05. The intraclass correlation coefficient (ICC) using the 2-way mixed effects model, absolute measurement, and single measurement were used to evaluate the interrater reliability of continuous data between the 2 groups [20], and this reliability was visualized using the Bland-Altman plot. The degree of agreement for all

variables was represented as a proportion. The accuracy of the chief concern and past medical history was classified into complete, partial, and fail. All statistical analyses were performed using R 3.6.0 (The R Foundation for Statistical Computing).

Results

During the study period, a total of 20,155 triage cases were processed at the hospital's ED, at an average of 194 cases per day. Among them, 1209 (6%) triage tasks were performed by the participants. After 146 cases were excluded by the criteria shown in Figure 2, a total of 1063 cases were used for study analysis.

The median time for participants to perform the triage task was 204 (IQR 155, 277) seconds with RMIS-AI and 231 (IQR 180, 313) seconds using manual input by EMR. The difference between the 2 methods was statistically significant (P<.001), as shown in the box plot in Figure 3.

The record completion rates of both methods for all triage cases are shown in Table 1. In the triage notes recorded by RMIS-AI, the first chief concern showed the highest record completion rate (81.84%), and all variables of vital signs that should be recorded as continuous variables showed comparable record completion rates of over 50% except for the respiratory rate. In most variables of the triage note, RMIS-AI showed a lower record completion rate than the manual method. However, in terms of recording additional chief concerns and past medical history, RMIS-AI showed a higher record completion rate than the manual method, which was statistically significant.

The accuracy of reproducing records by RMIS-AI for all variables is summarized in Table 2. In this study, only systolic blood pressure, diastolic blood pressure, oxygen saturation, and chief concern represented an accuracy of more than 50%, implying that RMIS-AI reproduced these variables recorded by the manual method by more than 50%. Furthermore, categorical variables such as past medical history and history of allergic episodes entered by RMIS-AI showed less accuracy than other variables.

Figure 4 shows the interrater reliability for continuous variables between the 2 methods. The ICC of systolic blood pressure and body temperature were 0.800 and 0.876, respectively, indicating substantial interrater reliability.

Figure 2. Flowchart of case inclusion. RMIS-AI: real-time medical record input assistance system with voice artificial intelligence.

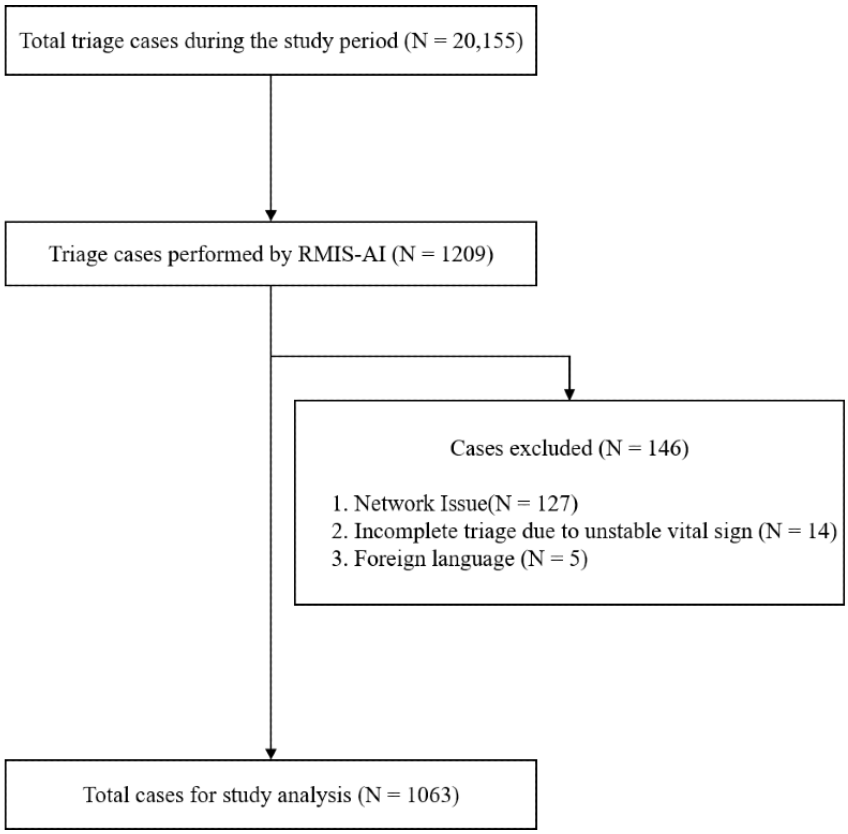


Figure 3. Comparison of median time for triage task, RMIS-AI (real-time medical record input assistance system with voice artificial intelligence) vs manual input.

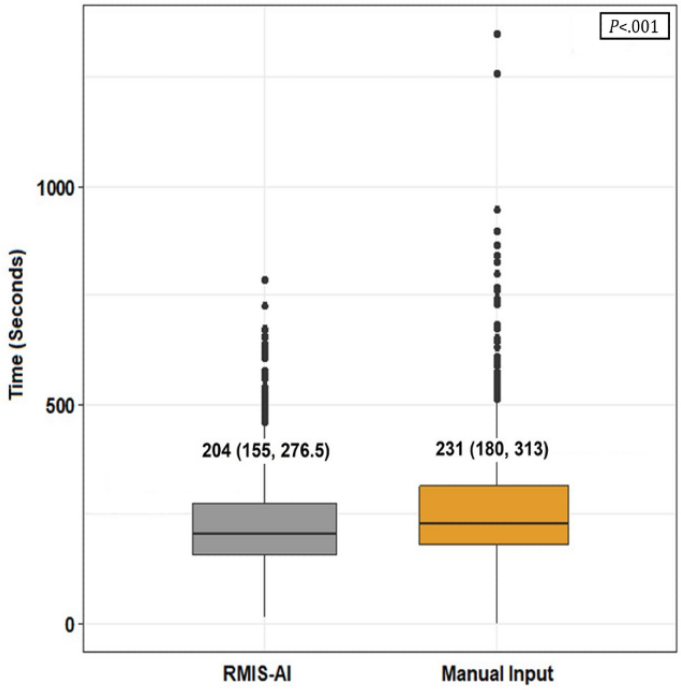


Table 1. Record completion rates of both methods.

Variable	Record completion cases, n (%)		*P* value
	RMIS-AI[a]	Manual input	
Chief concern, 1st	870 (81.84)	1063 (100)	<.001
Chief concern, 2nd	515 (48.45)	397 (37.35)	<.001
Chief concern, 3rd	230 (21.64)	106 (9.97)	<.001
History of allergic episode	257 (24.18)	1063 (100)	<.001
Past medical history, 1st	383 (36.03)	1030 (96.90)	<.001
Past medical history, 2nd	127 (11.95)	32 (3.01)	<.001
Past medical history, 3rd	27 (2.54)	12 (1.13)	.02
Systolic blood pressure	580 (54.56)	923 (86.83)	<.001
Diastolic blood pressure	578 (54.37)	923 (86.83)	<.001
Pulse rate	613 (57.67)	925 (87.02)	<.001
Respiratory rate	382 (35.94)	923 (86.83)	<.001
Body temperature	607 (57.10)	1061 (99.81)	<.001
Oxygen saturation	584 (54.94)	926 (87.11)	<.001

[a]RMIS-AI, real-time medical record input assistance system with voice artificial intelligence.

Table 2. Accuracy of RMIS-AI[a] compared to the manual method.

Variable	Cases with reproduction and cases with records by manual method, n/N (%)
Chief concern	
Complete reproduction[b]	366/1063 (34.43)
Partial reproduction[c]	190/1063 (17.87)
Failed reproduction[d]	507/1063 (49.41)
Past medical history	
Complete reproduction	226/1030 (21.94)
Partial reproduction	5/1030 (0.49)
Failed to reproduction	799/1080 (73.98)
History of allergic episode	158/1063 (14.68)
Systolic blood pressure	516/923 (55.90)
Diastolic blood pressure	495/923 (53.63)
Pulse rate	352/925 (38.05)
Respiratory rate	340/923 (36.84)
Body temperature	484/1061 (45.62)
Oxygen saturation	465/926 (50.22)

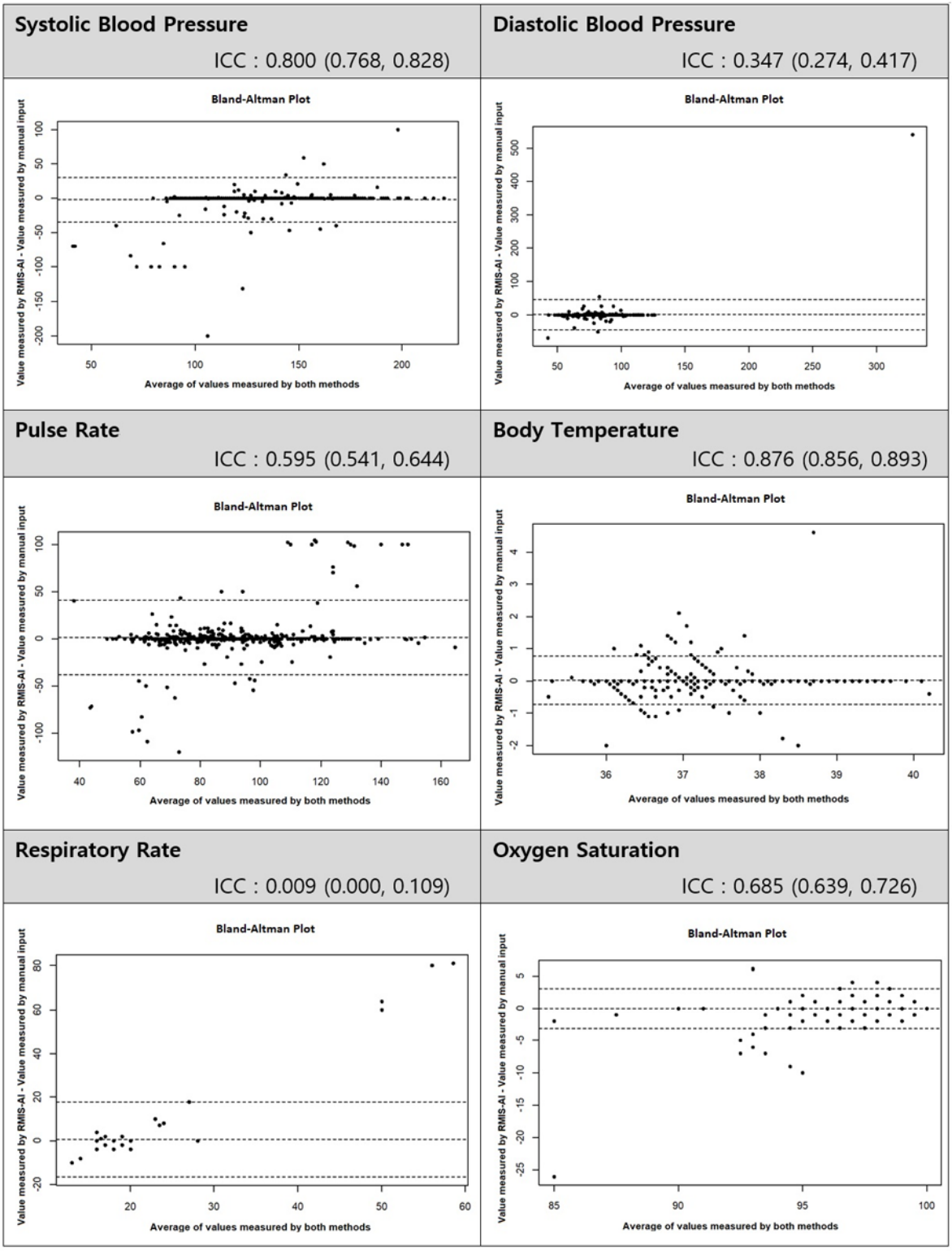
[a]RMIS-AI: real-time medical record input assistance system with voice artificial intelligence.

[b]All the values by manual input were reproduced by RMIS-AI.

[c]Partial values by manual input were reproduced by RMIS-AI.

[d]No values by manual input were reproduced by RMIS-AI.

Figure 4. Interrater reliability for continuous variables between 2 methods. ICC: intraclass correlation coefficient; RMIS-AI: real-time medical record input assistance system with voice artificial intelligence.



Discussion

Principal Findings

Previous study results have proven that prolonged waiting times and crowding are factors that reduce patient satisfaction and impair safety in the ED [21-25]. Long waiting times are known as the main cause of leaving without being seen after enrollment [26]. Leaving without being seen is considered an indicator of timeliness and effectiveness, which falls within the quality of care, as defined by the US Institute of Medicine, and poses a safety threat because it limits the options for patients to seek treatment elsewhere [27,28]. This study's results confirmed that the use of RMIS-AI in the ED shortens the time to perform the triage task. In cases where the patient influx of ED is rapidly increasing, reducing the time taken to perform the triage task could contribute to reducing patient waiting time. The triage

task mainly includes taking patient history through conversations and measuring vital signs. Typically, these actions and the recording of patient information are performed in separate steps. It is estimated that the RMIS-AI developed using voice recognition technology in our study reduced the time to perform the triage task by combining these tasks in a single step. After the COVID-19 outbreak, the screening process for patients visiting ED has been strengthened, resulting in a longer delay during the input phase [11]. In addition, medical staff who face the ED patients for the first time wear personal protective equipment to protect them from potential risk of infection [29], which made multitasking difficult for the medical staff of the triage unit. Therefore, the RMIS-AI developed using AI technology has proven its potential as a supportive solution to improve the quality of clinical practice in response to the new digital era as well as after the COVID-19 pandemic in the ED.

The record completion rates of RMIS-AI were inferior to the manual input by EMR in our study, especially in the input of allergy history or past medical history. In the case of categorical variables, such as allergy history or past medical history, NLP is more difficult than in the case of continuous variables, such as systolic blood pressure and pulse rate, because it is expressed in a wide variety of phrases rather than simple utterances. Korean is an agglutinative language and one of the morphologically rich and typologically diverse languages. Auxiliary, adverbial case markers, word spacing inconsistency, and the variety of expressions of predicates with the same meaning make NLP using Korean difficult. [30]. The engine was trained to input categorical variables, such as chief concern into 1 of 52 categories, and 13 categories for past medical history and binary format for allergy history. Because the sensitivity of the engine increases as there are more categories that can be input through NLP of the transcribed textual data, it was estimated that the record completion rate is low for past medical history and allergy history with relatively few categories. In addition, inferior results compared to manual input are observed presumably because triage nurses could not monitor the recording by RMIS-AI during triage tasks, and only recording by EMR was possible as usual. If the recording system of triage tasks using RMIS-AI compensates the conventional method, despite the time for triage task being longer than that reported in the study, it is expected that the record completion rate will be comparable to that of the manual method. For example, if triage nurses find that variables to be input by voice recognition have not been recorded during the task, they can speak to compensate for the missing variables. While recording patient history, it is common for triage performers to omit inessential information intentionally or forget acquired information. It has been reported that errors due to the inexperience of triage performers may adversely affect patients [4,5]. The recording by RMIS-AI involves relatively little subjectivity from the performer. Thus, RMIS-AI represents an alternative method that can offset the negative effects that occur because of the subjectivity of the triage performer. By recognizing various input values while recording patient history, it is possible to capture more detailed information that could not be detected using the conventional method. In this context, in the case of variables with multiple inputs, such as chief concern and past medical history, the record completion rate for subitems input by RMIS-AI was superior to that of the manual input.

In our study, it is assumed that the difference between the variables with and without relatively favorable accuracy is due to the complexity of NLP. NLP is still being developed as an artificial intelligence field, and because there is no standardized format, its performance is different depending on the type and amount of training data as well as the deep learning method applied [31,32]. For variables such as vital signs, the process ends with the triage nurse's voice passing through STT and charting the converted text numerically, but categorical variables such as chief concern should be categorized as textual data converted by STT into 1 of 52 categories through NLP. Variables of past medical history and history of allergic episode that the system was trained to classify into fewer categories had a lower record completion rate and failures for accurate production compared with the variables of chief concerns; this result is also presumed to be caused by the differences in NLP. In addition, the low record completion rate of RMIS-AI also led to inferior accuracy not being able to reproduce triage notes. In particular, the inferior accuracy of numerical variables applying the relatively uncomplicated NLP was attributed to the low record completion rate.

The reliability of pulse rate was lower than that of other vital sign values because there was a time difference between the input through RMIS-AI and the manual input because triage nurses record the pulse rate by watching the monitoring being measured as a continuous waveform. This result can be explained by the Bland-Altman plot, where the error range in the input value is narrow. In addition, the low ICC value of the respiratory rate was due to the less amount of data and low variability.

NLP is a tool that can structure unstructured textual data and enable the use of unstructured voice data that historically have not been used in the medical field. Previous studies have reported that the predictive performance of clinical outcomes is improved when unstructured textual data are used for machine learning in the medical field [13,15,16,33]. However, for unstructured textual data to be used in actual clinical practice rather than only in retrospective analyses, an environment in which STT is performed in real time should be developed. Most previous studies using textual data performed a retrospective analysis of text recorded on EMR using machine learning; therefore, they were not sufficient evidence in terms of improving clinical practice in the ED, which is a time-critical setting. Our study did not focus on prediction using a machine learning model with NLP but investigated the potential application of performing STT in a real clinical field through a prospective design. In this study, the machine learning framework was trained on unprocessed audio data. This approach can lead to an easy transition to a new system for acute clinical settings where decision-making should be efficient and precise in the digital era [19].

This study has several limitations. Although our study was conducted in a prospective design, a study using a randomized controlled design is needed to obtain definitive evidence that the RMIS-AI can replace the conventional method. Second, the

completeness and accuracy of the triage note by the current RMIS-AI are insufficient to safely replace the conventional manual input method. If NLP for the recording of triage note recording is learned using additional training material, it can be improved**.** Third, the reduction in time taken to perform the triage task does not guarantee improvements in patient outcomes. Therefore, the relationship between the use of RMIS-AI and improvement in clinical outcomes on patients in the ED should be investigated. Finally, the study was performed at an ED in a single tertiary hospital; thus, there is a limit to generalizing the research results.

Conclusions

In this study, we confirmed that the promptness in performing triage tasks improved using RMIS-AI developed with STT and NLP technology compared with the manual input method, but technical supplementation was required to deal with the current level of inferiority in sensitivity and accuracy. If similar studies are conducted to confirm the potential of such technologies in clinical practice, artificial intelligence could evolve as a supportive tool to improve patient experience.

Acknowledgments

The authors thank Medical Illustration and Design, part of the Medical Research Support Services of Yonsei University College of Medicine, for all artistic support related to this work. This study was supported by a faculty research grant of Yonsei University College of Medicine (6-2020-0117).

Conflicts of Interest

None declared.

References

1. Tang K, Ang C, Constantinides T, Rajinikanth V, Acharya U, Cheong K. Artificial Intelligence and Machine Learning in Emergency Medicine. Biocybernetics and Biomedical Engineering 2021 Jan;41(1):156-172 [FREE Full text] [doi: 10.1016/j.bbe.2020.12.002]
2. Christ M, Grossmann F, Winter D, Bingisser R, Platz E. Modern triage in the emergency department. Dtsch Arztebl Int 2010 Dec;107(50):892-898 [FREE Full text] [doi: 10.3238/arztebl.2010.0892] [Medline: 21246025]
3. Jarvis PRE. Improving emergency department patient flow. Clin Exp Emerg Med 2016 Jun;3(2):63-68 [FREE Full text] [doi: 10.15441/ceem.16.127] [Medline: 27752619]
4. Dugas AF, Kirsch TD, Toerper M, Korley F, Yenokyan G, France D, et al. An Electronic Emergency Triage System to Improve Patient Distribution by Critical Outcomes. J Emerg Med 2016 Jun;50(6):910-918. [doi: 10.1016/j.jemermed.2016.02.026] [Medline: 27133736]
5. Lee JH, Park YS, Park IC, Lee HS, Kim JH, Park JM, et al. Over-triage occurs when considering the patient's pain in Korean Triage and Acuity Scale (KTAS). PLoS One 2019 Dec;14(5):e0216519-e0216800 [FREE Full text] [doi: 10.1371/journal.pone.0216519] [Medline: 31071132]
6. Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A, et al. Predicting Intensive Care Unit admission among patients presenting to the emergency department using machine learning and natural language processing. PLoS One 2020;15(3):e0229331 [FREE Full text] [doi: 10.1371/journal.pone.0229331] [Medline: 32126097]
7. Kwon JM, Lee Y, Lee Y, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. PLoS One 2018;13(10):e0205836 [FREE Full text] [doi: 10.1371/journal.pone.0205836] [Medline: 30321231]
8. Yu JY, Jeong GY, Jeong OS, Chang DK, Cha WC. Machine Learning and Initial Nursing Assessment-Based Triage System for Emergency Department. Healthc Inform Res 2020 Jan;26(1):13-19 [FREE Full text] [doi: 10.4258/hir.2020.26.1.13] [Medline: 32082696]
9. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019 Feb 22;23(1):64 [FREE Full text] [doi: 10.1186/s13054-019-2351-7] [Medline: 30795786]
10. Goto T, Camargo CA, Faridi MK, Freishtat RJ, Hasegawa K. Machine Learning-Based Prediction of Clinical Outcomes for Children During Emergency Department Triage. JAMA Netw Open 2019 Jan 04;2(1):e186937 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.6937] [Medline: 30646206]
11. Wang J, Zong L, Zhang J, Sun H, Harold Walline J, Sun P, et al. Identifying the effects of an upgraded 'fever clinic' on COVID-19 control and the workload of emergency department: retrospective study in a tertiary hospital in China. BMJ Open 2020 Aug 20;10(8):e039177 [FREE Full text] [doi: 10.1136/bmjopen-2020-039177] [Medline: 32819955]
12. Lee S, Mohr NM, Street WN, Nadkarni P. Machine Learning in Relation to Emergency Medicine Clinical and Operational Scenarios: An Overview. West J Emerg Med 2019 Mar;20(2):219-227 [FREE Full text] [doi: 10.5811/westjem.2019.1.41244] [Medline: 30881539]

13. Roquette BP, Nagano H, Marujo EC, Maiorano AC. Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. Neural Netw 2020 Jun;126:170-177. [doi: 10.1016/j.neunet.2020.03.012] [Medline: 32240912]
14. Sterling NW, Patzer RE, Di M, Schrager JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. Int J Med Inform 2019 Sep;129:184-188. [doi: 10.1016/j.ijmedinf.2019.06.008] [Medline: 31445253]
15. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager JD. Prediction of Emergency Department Hospital Admission Based on Natural Language Processing and Neural Networks. Methods Inf Med 2017 Oct 26;56(5):377-389. [doi: 10.3414/ME17-01-0024] [Medline: 28816338]
16. Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A, et al. Risk of mortality and cardiopulmonary arrest in critical patients presenting to the emergency department using machine learning and natural language processing. PLoS One 2020;15(4):e0230876 [FREE Full text] [doi: 10.1371/journal.pone.0230876] [Medline: 32240233]
17. Tootooni MS, Pasupathy KS, Heaton HA, Clements CM, Sir MY. CCMapper: An adaptive NLP-based free-text chief complaint mapping algorithm. Comput Biol Med 2019 Oct;113:103398. [doi: 10.1016/j.compbiomed.2019.103398] [Medline: 31454613]
18. Frost DW, Vembu S, Wang J, Tu K, Morris Q, Abrams HB. Using the Electronic Medical Record to Identify Patients at High Risk for Frequent Emergency Department Visits and High System Costs. Am J Med 2017 May;130(5):601.e17-601.e22. [doi: 10.1016/j.amjmed.2016.12.008] [Medline: 28065773]
19. Blomberg SN, Folke F, Ersbøll AK, Christensen HC, Torp-Pedersen C, Sayre MR, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation 2019 May;138:322-329 [FREE Full text] [doi: 10.1016/j.resuscitation.2019.01.015] [Medline: 30664917]
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
21. Hunt MT, Glucksman ME. A review of 7 years of complaints in an inner-city accident and emergency department. Arch Emerg Med 1991 Mar;8(1):17-23 [FREE Full text] [doi: 10.1136/emj.8.1.17] [Medline: 1854388]
22. Vezyridis P, Timmons S. National targets, process transformation and local consequences in an NHS emergency department (ED): a qualitative study. BMC Emerg Med 2014 Jun 13;14:12 [FREE Full text] [doi: 10.1186/1471-227X-14-12] [Medline: 24927819]
23. Jo S, Jeong T, Jin YH, Lee JB, Yoon J, Park B. ED crowding is associated with inpatient mortality among critically ill patients admitted via the ED: post hoc analysis from a retrospective study. Am J Emerg Med 2015 Dec;33(12):1725-1731. [doi: 10.1016/j.ajem.2015.08.004] [Medline: 26336833]
24. Singer AJ, Thode HC, Viccellio P, Pines JM. The association between length of emergency department boarding and mortality. Acad Emerg Med 2011 Dec;18(12):1324-1329 [FREE Full text] [doi: 10.1111/j.1553-2712.2011.01236.x] [Medline: 22168198]
25. Sprivulis PC, Da Silva JA, Jacobs IG, Frazer ARL, Jelinek GA. The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. Med J Aust 2006 Mar 06;184(5):208-212. [doi: 10.5694/j.1326-5377.2006.tb00416.x] [Medline: 16515429]
26. Shaikh SB, Jerrard DA, Witting MD, Winters ME, Brodeur MN. How long are patients willing to wait in the emergency department before leaving without being seen? West J Emerg Med 2012 Dec;13(6):463-467 [FREE Full text] [doi: 10.5811/westjem.2012.3.6895] [Medline: 23359833]
27. Sheraton M, Gooch C, Kashyap R. Patients leaving without being seen from the emergency department: A prediction model using machine learning on a nationwide database. J Am Coll Emerg Physicians Open 2020 Dec;1(6):1684-1690 [FREE Full text] [doi: 10.1002/emp2.12266] [Medline: 33392577]
28. Welch SJ, Asplin BR, Stone-Griffith S, Davidson SJ, Augustine J, Schuur J, Emergency Department Benchmarking Alliance. Emergency department operational metrics, measures and definitions: results of the Second Performance Measures and Benchmarking Summit. Ann Emerg Med 2011 Jul;58(1):33-40. [doi: 10.1016/j.annemergmed.2010.08.040] [Medline: 21067846]
29. Lee SJ, Choi A, Ryoo HW, Pak YS, Kim HC, Kim JH. Changes in Clinical Characteristics among Febrile Patients Visiting the Emergency Department before and after the COVID-19 Outbreak. Yonsei Med J 2021 Dec;62(12):1136-1144 [FREE Full text] [doi: 10.3349/ymj.2021.62.12.1136] [Medline: 34816644]
30. Shin D, Kam HJ, Jeon M, Kim HY. Automatic Classification of Thyroid Findings Using Static and Contextualized Ensemble Natural Language Processing Systems: Development Study. JMIR Med Inform 2021 Sep 21;9(9):e30223 [FREE Full text] [doi: 10.2196/30223] [Medline: 34546183]
31. Luo JW, Chong JJR. Review of Natural Language Processing in Radiology. Neuroimaging Clin N Am 2020 Nov;30(4):447-458. [doi: 10.1016/j.nic.2020.08.001] [Medline: 33038995]
32. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017 Sep;73:14-29 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]

33. Chen CH, Hsieh JG, Cheng SJ, Lin YL, Lin PH, Jeng JH. Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. Int J Med Inform 2020 Jul;139:104146. [doi: 10.1016/j.ijmedinf.2020.104146] [Medline: 32387818]

Abbreviations

ED: emergency department
EMR: electronic medical record
ICC: intraclass correlation coefficient
NLP: natural language processing
RMIS-AI: real-time medical record input assistance system with voice artificial intelligence
STT: speech-to-text

Edited by C Lovis; submitted 27.05.22; peer-reviewed by S Blomberg, M Bullard; comments to author 03.07.22; revised version received 27.07.22; accepted 15.08.22; published 31.08.22.

Please cite as:
Cho A, Min IK, Hong S, Chung HS, Lee HS, Kim JH
Effect of Applying a Real-Time Medical Record Input Assistance System With Voice Artificial Intelligence on Triage Task Performance in the Emergency Department: Prospective Interventional Study
JMIR Med Inform 2022;10(8):e39892
URL: https://medinform.jmir.org/2022/8/e39892
doi:10.2196/39892
PMID:36044254

Corrigenda and Addenda

Addendum: Building a Shared, Scalable, and Sustainable Source for the Problem-Oriented Medical Record: Developmental Study

Christophe Gaudet-Blavignac[1,2], BSc, MSc; Andrea Rudaz[3], MHSA, MD; Christian Lovis[1,2], MPH, MD

[1]Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland
[2]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
[3]Medical and Quality Directorate, Geneva University Hospitals, Geneva, Switzerland

Corresponding Author:
Christophe Gaudet-Blavignac, BSc, MSc
Division of Medical Information Sciences
Geneva University Hospitals
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 223726201
Email: christophe.gaudet-blavignac@hcuge.ch

Related Article:

Correction of: https://medinform.jmir.org/2021/10/e29174

(JMIR Med Inform 2022;10(8):e41257) doi:10.2196/41257

In "Building a Shared, Scalable, and Sustainable Source for the Problem-Oriented Medical Record: Developmental Study" (JMIR Med Inform 2021;9(10):e29174), the authors made two updates.

After the publication of the original article, the Geneva University Hospitals Common Problem List was released to the public under the Creative Commons CC BY-SA 4.0 license. It was important to the authors that the readers of the article knew that the list was available for reuse.

1. Accordingly, the following sentence was added at the end of the *Acknowledgments* section:

> *The common problem list is available under the Creative Commons CC BY-SA 4.0 license on Yareta, the digital solution of the University of Geneva for archiving and preserving research data [40].*

2. Full citation of this new reference [40] has been added to the article's *References* section.

The correction will appear in the online version of the paper on the JMIR Publications website on August 9, 2022, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

40. Gaudet-Blavignac C. Geneva University Hospitals Common Problem List. Yareta. URL: https://doi.org/10.26037/YARETA:NAEGEJQVXZFWLIU236PXN5LUS4 [accessed 2022-07-20]

Submitted 20.07.22; this is a non–peer-reviewed article;accepted 21.07.22; published 09.08.22.

Please cite as:
Gaudet-Blavignac C, Rudaz A, Lovis C
Addendum: Building a Shared, Scalable, and Sustainable Source for the Problem-Oriented Medical Record: Developmental Study
JMIR Med Inform 2022;10(8):e41257
URL: https://medinform.jmir.org/2022/8/e41257
doi:10.2196/41257
PMID:35944251

Corrigenda and Addenda

Correction: The Science of Learning Health Systems: Scoping Review of Empirical Research

Louise A Ellis[1], PhD; Mitchell Sarkies[1], PhD; Kate Churruca[1], PhD; Genevieve Dammery[1], BSc (Hons); Isabelle Meulenbroeks[1], MRes; Carolynn L Smith[1], PhD; Chiara Pomare[1], PhD; Zeyad Mahmoud[1], PhD; Yvonne Zurynski[1], PhD; Jeffrey Braithwaite[1], PhD

Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

Corresponding Author:
Louise A Ellis, PhD
Australian Institute of Health Innovation
Macquarie University
75 Talavera Rd
Sydney, 2113
Australia
Phone: 61 298502484
Fax: 61 298502499
Email: louise.ellis@mq.edu.au

Related Article:

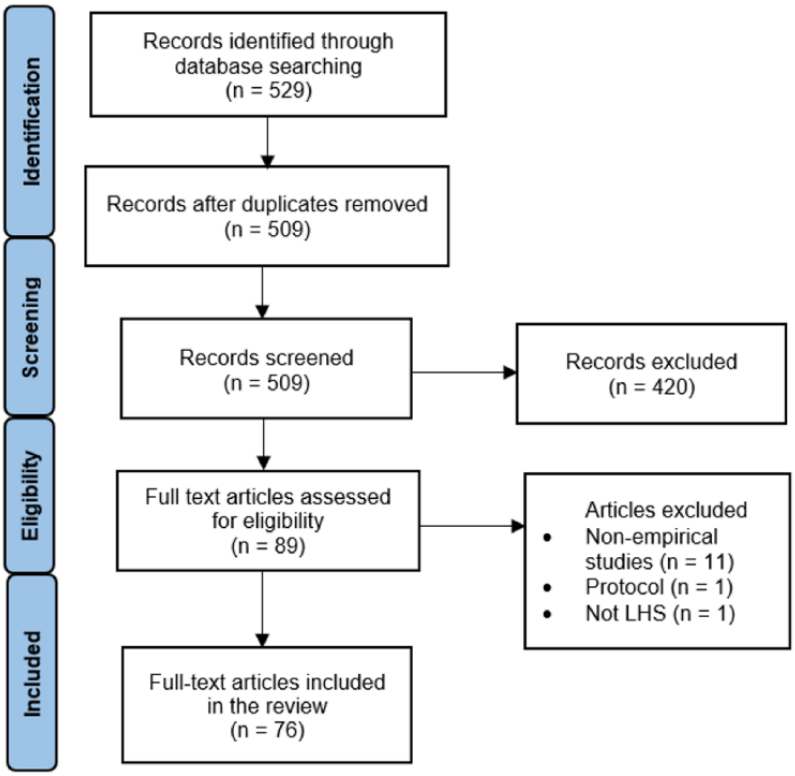
Correction of: https://medinform.jmir.org/2022/2/e34907

(JMIR Med Inform 2022;10(8):e41424) doi:10.2196/41424

In "The Science of Learning Health Systems: Scoping Review of Empirical Research" (2022;10(2):e34907) the authors noted an error.

In the originally published article, Figure 2 appeared incorrectly (Multimedia Appendix 1). In the corrected version of the article, Figure 2 was updated with the following image:

Figure 2. Search and review strategy. LHS: learning health system.



The correction will appear in the online version of the paper on the JMIR Publications website on August 4, 2022, together with the publication of this correction notice. Because this was made after submission to full-text repositories, the corrected article has also been resubmitted to those repositories.

Multimedia Appendix 1
Originally published Figure 1.
[PNG File , 13 KB - medinform_v10i8e41424_app1.png]

Submitted 25.07.22; this is a non–peer-reviewed article;accepted 29.07.22; published 04.08.22.

Please cite as:
Ellis LA, Sarkies M, Churruca K, Dammery G, Meulenbroeks I, Smith CL, Pomare C, Mahmoud Z, Zurynski Y, Braithwaite J
Correction: The Science of Learning Health Systems: Scoping Review of Empirical Research
JMIR Med Inform 2022;10(8):e41424
URL: https://medinform.jmir.org/2022/8/e41424
doi:10.2196/41424
PMID:35926194

Original Paper

Deployment of a Free-Text Analytics Platform at a UK National Health Service Research Hospital: CogStack at University College London Hospitals

Kawsar Noor[1,2,3,4]; Lukasz Roguski[1,2,3]; Xi Bai[1,2,3]; Alex Handy[1,2,3,4]; Roman Klapaukh[4]; Amos Folarin[1,2,4,5,6]; Luis Romao[2,3,4]; Joshua Matteson[7]; Nathan Lea[2,3,4]; Leilei Zhu[3]; Folkert W Asselbergs[2,3]; Wai Keong Wong[3]; Anoop Shah[1,2,3,4]; Richard JB Dobson[1,2,3,4,5,6]

[1]University College London, London, United Kingdom

[2]Institute of Health Informatics, University College London, London, United Kingdom

[3]National Institute for Health and Care Research Biomedical Research Centre, University College London Hospitals National Health Service Foundation Trust, London, United Kingdom

[4]Health Data Research UK London, University College London, London, United Kingdom

[5]National Institute for Health and Care Research Biomedical Research Centre, South London and Maudsley National Health Service Foundation Trust, King's College London, London, United Kingdom

[6]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[7]Epic Systems Corporation, London, United Kingdom

Corresponding Author:
Kawsar Noor
University College London
222 Euston Rd
London, NW1 2DA
United Kingdom
Phone: 44 7837262856
Email: kawsar.noor.15@ucl.ac.uk

Abstract

Background: As more health care organizations transition to using electronic health record (EHR) systems, it is important for these organizations to maximize the secondary use of their data to support service improvement and clinical research. These organizations will find it challenging to have systems capable of harnessing the unstructured data fields in the record (clinical notes, letters, etc) and more practically have such systems interact with all of the hospital data systems (legacy and current).

Objective: We describe the deployment of the EHR interfacing information extraction and retrieval platform CogStack at University College London Hospitals (UCLH).

Methods: At UCLH, we have deployed the CogStack platform, an information retrieval platform with natural language processing capabilities. The platform addresses the problem of data ingestion and harmonization from multiple data sources using the Apache NiFi module for managing complex data flows. The platform also facilitates the extraction of structured data from free-text records through use of the MedCAT natural language processing library. Finally, data science tools are made available to support data scientists and the development of downstream applications dependent upon data ingested and analyzed by CogStack.

Results: The platform has been deployed at the hospital, and in particular, it has facilitated a number of research and service evaluation projects. To date, we have processed over 30 million records, and the insights produced from CogStack have informed a number of clinical research use cases at the hospital.

Conclusions: The CogStack platform can be configured to handle the data ingestion and harmonization challenges faced by a hospital. More importantly, the platform enables the hospital to unlock important clinical information from the unstructured portion of the record using natural language processing technology.

(JMIR Med Inform 2022;10(8):e38122) doi:10.2196/38122

KEYWORDS

natural language processing; text mining; information retrieval; electronic health record system; clinical support

Introduction

Background

Over the past 20 years, we have seen an increased uptake of electronic health records (EHRs) within health care organizations, with much of this being attributable to national efforts in having health care organizations transition to using full EHR systems [1,2]. These EHRs represent a rich data asset, but there remains a challenge in the secondary use of the data for improving clinical care through activities, such as service improvement and clinical research. In many cases, EHRs have simply replicated the paper system that they replaced and have not taken full advantage of the opportunities presented in having the health records in this new electronic format. While functional systems to address these gaps are emerging, many of the tools and data analytic approaches used on EHR data are limited to structured data, such as coded diagnoses and numeric clinical measurements. However, the structured data only account for a small portion of the EHR data, as it is estimated that almost 80% of information records remain unstructured in the form of images, free-text records, and other such unstructured data formats [3]. In particular, the free-text records often contain important clinical information, such as patient diagnoses, that have not yet been recorded as structured data [4]. An additional difficulty is that a hospital's record is typically distributed across numerous disconnected data systems, which presents a challenge in data harmonization.

Working with EHRs thus presents challenges firstly in harmonizing and accessing the hospitals entire record from both existing and legacy data systems and secondly having tools and techniques available to mine and extract data from within these records, especially the unstructured free text. Manual analysis of unstructured text is time-consuming, so there has been much interest in developing automated methods for extracting accurate structured information from the free-text records [5]. Interpreting free text is a major analytic challenge; clinical text is written in a variety of styles by numerous authors and may have misspellings, negations, and other linguistic features. There has been intense interest in developing natural language processing (NLP) techniques to interpret clinical text [6,7]. Early methods used a rule-based approach, but more modern algorithms incorporate machine learning techniques, enabling the algorithms to "learn" as more data are analyzed.

The CogStack platform [8] was developed to address these exact problems. The platform can be described as an information retrieval system designed to interface with a hospital's EHR system. It was initially developed with an emphasis on ingestion and harmonization of records from multiple data systems within a health care organization. While certain off-the-shelf NLP tools were explored in the first iteration, they were added as a proof of concept to demonstrate that the platform could potentially be configured to interact with such tools.

In this paper, we discuss the experience of deploying CogStack at University College London Hospitals (UCLH) and highlight modifications to the platform that have improved its data harmonization and NLP capabilities. Our deployment of CogStack has focused on addressing the following 3 key issues that we feel are universal to all research driven health care organizations.

Multiple Data Systems

The EHRs of an organization will typically be distributed across a number of different vendor systems, posing a challenge for the use of this information for clinical care and research. It is not uncommon for an organization to have to maintain oversight over a myriad of data systems and vendors due to the fact that different clinical specialties will have different requirements of how data needs to be stored and managed. The resulting heterogeneity in data means that it is challenging for the organization to find a common data model or even process through which the organization's entire record can be harmonized. Methods and systems through which data are stored, collected, and retrieved have been improving in order to tackle this challenge. Most notably, many National Health Service (NHS) trusts have opted to transition to using full-scale EHR systems (eg, Epic), each of which typically enforce their own data models. Some systems, such as Epic, go further in providing additional systems that allow data from third-party data and legacy systems to be integrated with data collected via their own systems (Epic Clarity/Caboodle). Messaging standards (eg, HL7 Fast Healthcare Interoperability Resources [FHIR] [9]), standardized terminologies (eg, Systematized Nomenclature of Medicine -- Clinical Terms [SNOMED CT]), and standardized clinical information models (eg, openEHR archetypes [10]) aim to improve interoperability between systems, but much more work is needed in this area. In order to maximize the benefit of patient data, it is essential that clinicians and researchers can access data in a way that is flexible, easily adaptable, and independent of the organization's choice of current and previous EHR systems.

Multiple Data Formats

A patient's record may be distributed across both scanned documents (PDFs) and text documents (.doc files), and data may be stored in relational databases. Legacy documents, for example, will likely be stored as files and attachments, whereas data that have been generated using a modern EHR system will likely be stored in a more structured way, possibly in a relational database. An information retrieval system would thus need to be able to ingest and interact with records from all the various data formats used by the organization. The CogStack platform provides functionality for document processing, including PDF to text conversion, or optical character recognition that may be needed prior to analysis of the text itself.

Unstructured Text

A final issue is that data within the EHR systems are recorded in both structured and unstructured fields. Some information is inherently unstructured in nature and needs to be recorded as free text (eg, patient stories), but even where structured fields

are available, clinicians may not use them and enter the information in free text instead. For example, a recent audit in our trust found that patients admitted with suspected or confirmed COVID-19 had only 62.3% of their key diagnoses and comorbidities recorded in the structured problem list [4]. In order to support use of clinical data at scale and for multiple stakeholders, a successful information retrieval system should provide mechanisms through which the clinical information within the unstructured free-text notes can be made available. The CogStack platform provides a convenient user interface for searching free text, invoking information extraction algorithms, and presenting the results in a way that is easy to visualize and harness for downstream research or for reintegration as structured data back into the EHR.

There has been a great deal of interest in integrating NLP systems with EHRs to tackle the problem of unlocking value from unstructured data [11]. A number of commercial vendors have proposed NLP analysis as a service, where the vendor supplies NLP models that are used to process unstructured data [12-14]. In general, to our understanding, the NLP engines used by these vendors are trained using nontrust data and are generally not easily fine-tuned. In contrast, CogStack is a fully open-source platform, and the underlying NLP technology is tuned using the hospital's data and deployed on hospital infrastructure. Furthermore, the intellectual property for the NLP engines is not owned by the vendor and instead is proprietary to the hospital.

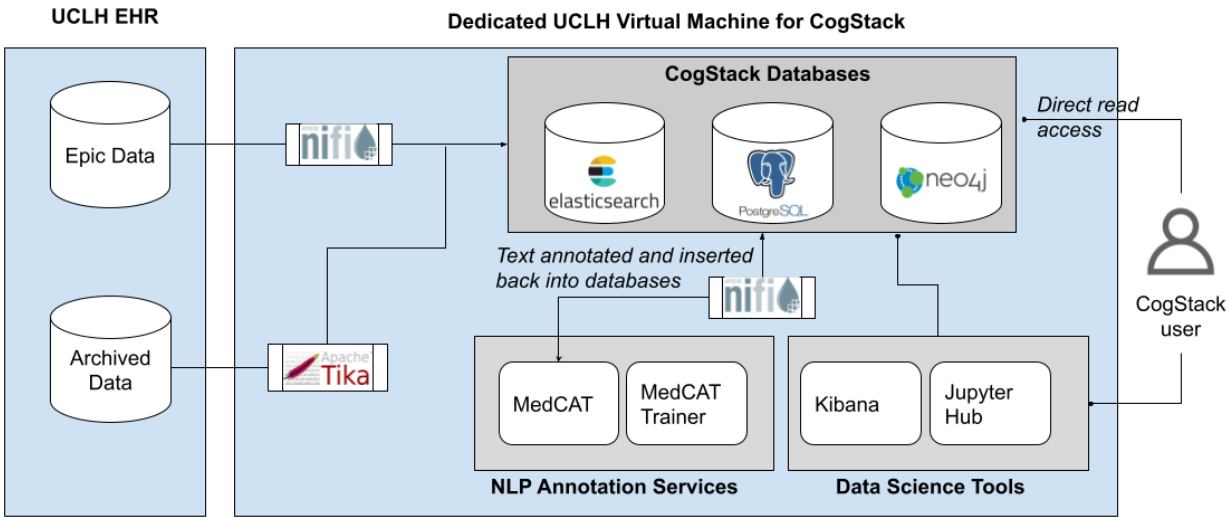
In the rest of the paper, we describe the deployment of CogStack at UCLH and demonstrate how it has been configured to handle commonly seen use cases within the hospital. In the Results section, we demonstrate that it has been or is being currently used to support several service evaluation and research projects within the hospital.

Methods

Overview

In this section, we describe the various components of the CogStack platform [15] and describe how the platform has been deployed and configured at UCLH. Figure 1 depicts the various components and how they have been configured at UCLH. Broadly speaking, the platform provides 3 categories of functionality, namely, the ability to read data from the hospital's EHR system, to store data, and to interact with the data programmatically on various NLP tools and interfaces.

Figure 1. An overview of the CogStack platform as deployed at University College London Hospitals (UCLH). EHR: electronic health record; NLP: natural language processing.



Infrastructure

The CogStack platform deployed at UCLH builds upon the previous version [8] that has been deployed at multiple hospitals, including South London and Maudsley Hospital, Guy's and St Thomas' Hospital, and King's College Hospital. In particular, the latest version provides 2 key updates. The first is related to the improvements in the platform's NLP capabilities, and the second relates to the use of Apache NiFi to manage the various data flows within the platform.

The first update is the use of the MedCAT NLP toolkit to provide clinical concept detection capabilities. The MedCAT tool is used to detect and extract clinical information from the free-text records (diagnosis, procedures, etc). The second update is the use of Apache NiFi for managing data flows within the platform. This was added based on lessons learnt from the previous iteration of CogStack, where the platform required the development of a number of custom extract, transform, and load (ETL) scripts for managing the ingestion of data from the live record and legacy systems. This approach however does not scale well in practice, and it can quickly become burdensome for developers to manage the various ETL scripts when the number of data flows increases. Observing these difficulties, the Apache NiFi module was added to the CogStack platform. Apache NiFi is a visual interface for managing complex data flows between different data systems. Data flows in Apache NiFi are depicted as directed graphs and provide useful visual feedback for system administrators, such as the status of a particular data flow and the number of documents processed. Most importantly, Apache NiFi is compatible with various data systems, which means that administrators are capable of writing

the various ETL components in practically whatever programming language they choose. In addition, all of the data flows are accessible within a single interface, and this makes maintaining oversight of all of the data flows considerably easier than having to monitor multiple custom ETL scripts. UCLH has developed a number of NiFi workflows that are designed to work with the UCLH data warehouse as well as legacy data systems. These NiFi components conduct the various extractions and data transformations necessary for downstream CogStack services. We discuss these various data flows in the next section.

Data Security and Governance

Use of unstructured EHR data for clinical research is challenging because of confidentiality concerns, leading to difficulty in obtaining ethics and information governance approvals for accessing such data. The CogStack approach is to embed text analytic capabilities and research staff within NHS Trusts, allowing sensitive text to be analyzed in situ.

Although data are routinely ingested into the CogStack platform, researchers wishing to use the data still need to undergo an approval process before accessing the data or making use of machine learning models trained on patient data for their research. UCLH has in place a system called Data Explorer [16] through which researchers can apply for access to use clinical data. If researchers require CogStack, an application needs to be submitted through the Data Explorer system and approved, and the appropriate data protection impact assessments (DPIAs) need to be completed. Each DPIA is assessed and approved by UCLH's information governance lead before the user is able to access the data on the CogStack platform, and eventually, the permission to process and analyze the data using CogStack-trained machine learning models is provided.

As data ingested into the CogStack platform involve patient-sensitive information, all UCLH CogStack services are hosted within a secure environment that is only accessible within the hospital network. CogStack has a number of virtual machines that have been provisioned to process the trust's data. We have followed the best practice for software deployment and have designated these virtual machines for development, testing, and production.

In addition, we have in place processes to be able to remove patient-identifiable data from the free-text records before use for research. CogStack has a deidentification module that is used to prepare batches of data for specific users and can be deployed before or after ingestion into CogStack's central standardized data lake. The module builds on the open-source Philter library developed by the University of California, San Francisco, which achieved over 99% recall on the benchmark I2B2 deidentification data set by using a combination of rule-based and statistical approaches [17]. In the following text, we detail the ingestion pipeline as well as how data are accessed and processed once ingested into the platform.

Data Ingestion

CogStack uses Apache NiFi for managing data flows from the hospital's various data sources into CogStack's databases. In Figure 1, data flows can be seen between the live EHR (Epic data) and the hospital's archived data warehouse. Table 1 provides a summary of the number of documents ingested so far into the platform. Using Apache NiFi, we are able to define how the ETL processes are implemented for each data source. We are able to set up data flows that run periodically, as well as manage ingestions that only happen once. Below, we describe the various data sources from which we ingest data.

Table 1. Number of notes ingested and analyzed by CogStack.

Document type	Number of notes
Clinical notes	10,500,000
Imaging reports	2,000,000
Clinical letters	3,000,000
Archived records	16,000,000

Trust Data

In 2019, UCLH officially transitioned to using Epic [18] as its primary EHR system. Prior to this, the trust had a number of data systems for each of its departments/clinics. The Epic system has in place a number of databases that capture integrated hospital data. Its data warehouse, Caboodle, has been extended to capture non-EHR and historic data records as well. UCLH has deployed the Epic Caboodle data warehouse for this purpose, and this is the primary database that the CogStack platform ingests data from.

As these data are stored as relational data, setting up data flows into CogStack requires only that CogStack understands the data schema of the target database. The data flows are then set up using Apache NiFi as a batch process. The batch process transforms the data into a format that is compatible with the various CogStack databases. Most clinical research projects requiring CogStack to date have been retrospective studies and have not required access to a live data feed. Consequently, the batch process runs on a daily basis, and this can be easily modified as needed through the NiFi interface.

Archived Data and Other Records

A number of records in the trust (such as those created prior to the transition to Epic) are not included in the Epic Caboodle data feed and require custom data flows to be set up. Records in the legacy systems are often stored as documents that have been scanned as images or as text documents (eg, .doc files, .pdf files, etc). In such cases, CogStack uses Apache Tika's optical character recognition software to convert the contents of these documents into text that can then be saved into the platform's various databases.

One-Off Ingestions

While CogStack's primary focus is ingesting and processing data from the trust, there are occasionally requests to analyze nontrust data sets. Examples of this include allergy reports taken from the National Reporting and Learning System. In such cases, CogStack can accommodate these ad hoc requests via custom ingestion scripts using Apache NiFi.

Data Storage

As can be seen in Figure 1, the CogStack platform at UCLH saves its ingested data into 3 types of databases. This is to cater to the needs of the different types of users/downstream consumers of CogStack data. Once ingested, data can subsequently be accessed directly via a read-only user account or by using the set of data science tools that CogStack provides.

The first database provided is the ElasticSearch database, which is particularly useful for users and applications working with free-text data owing to its text-based indexing and querying capabilities. The second database is a PostGres database, which allows relational modeling of data and is more importantly widely compatible for many downstream users. Lastly, there has been recent work in ingesting data into a Ne04j database. This is to support the storage of graph-like data structures (eg, SNOMED ontology relations).

NLP Services

The core NLP functionality of the platform is provided by the MedCAT NLP toolkit [19]. The MedCAT toolkit is a named entity recognition and linking model that can identify clinical concepts in free text and link them to a predefined medical ontology (eg, SNOMED CT and UMLS). Currently, a UCLH-trained MedCAT model is deployed as a RESTful application programming interface (API) service and is scheduled via the Apache NiFi module to batch annotate new documents that have been inserted into the CogStack databases.

The underlying approach used by MedCAT is dependent on a neural network–based approach that learns latent representations (concept embeddings) of clinical concepts based on how they appear in free text. The underlying algorithm is a modified version of the word2vec algorithm, which learns numerical representations of a word based on the words that surround it.

Training MedCAT is done in 2 phases. The first phase is a self-supervised phase in which MedCAT employs a simple technique to preannotate a large corpus of clinical text. In this step, the algorithm identifies string matches for each concept synonym in the medical ontology being used (eg, searching for matches of "lung cancer" in each document). Once identified, the word2vec algorithm is used to learn embeddings for those identified entities within the documents. This process provides MedCAT with an initial representation for how the concepts are represented in free text.

In the second phase, the model is fine-tuned using human-provided annotations. In this case, the model is taught to predict the correct label as provided by the human annotator using the MedCAT trainer interface. Based on some previous studies [19], the number of annotations required for fine-tuning is small (500-600 annotated documents).

Collecting annotated data for training machine learning models is done through a custom annotation interface. A custom interface was chosen over off-the-shelf ones (eg, Doccano) as many of our annotation use cases require integrated tools for searching for clinical information.

MedCAT is trained using the MedCAT trainer interface [20]. The interface allows a user to load documents to be annotated by multiple annotators. The interface also provides an active learning mode that enables generated annotations to be used to retrain an existing MedCAT model in real time. The performance of the model can also be tracked in real time so the users can monitor performance change with additional annotations.

In addition to identifying clinical concepts in text, MedCAT provides a wrapper for training additional machine learning models for identifying important meta information for the extracted entities. Meta information of interest may include entity negated (eg, "patient does not have fever symptoms"), if an identified entity relates to the patient or to somebody else (the experiencer), or whether it is current or historic. In order to implement these models, MedCAT uses a sequence-based classifier (Bi-LSTMs) that takes the surrounding words of the identified terms and trains a classifier to predict if the meta label is assignable or not.

As mentioned earlier, at present, MedCAT is used to annotate documents that have been ingested into the platform. The annotations are saved in all 3 databases to ensure the end users have the ability to query whatever database they wish to use. The MedCAT models are trained using unsupervised learning based on records ingested into the platform. The model is occasionally fine-tuned when clinicians submit annotations via the MedCAT trainer interface. It is also useful to note that MedCAT models have been shown to generalize well across multiple hospital settings with only minimal fine-tuning required [19].

Data Science Tools

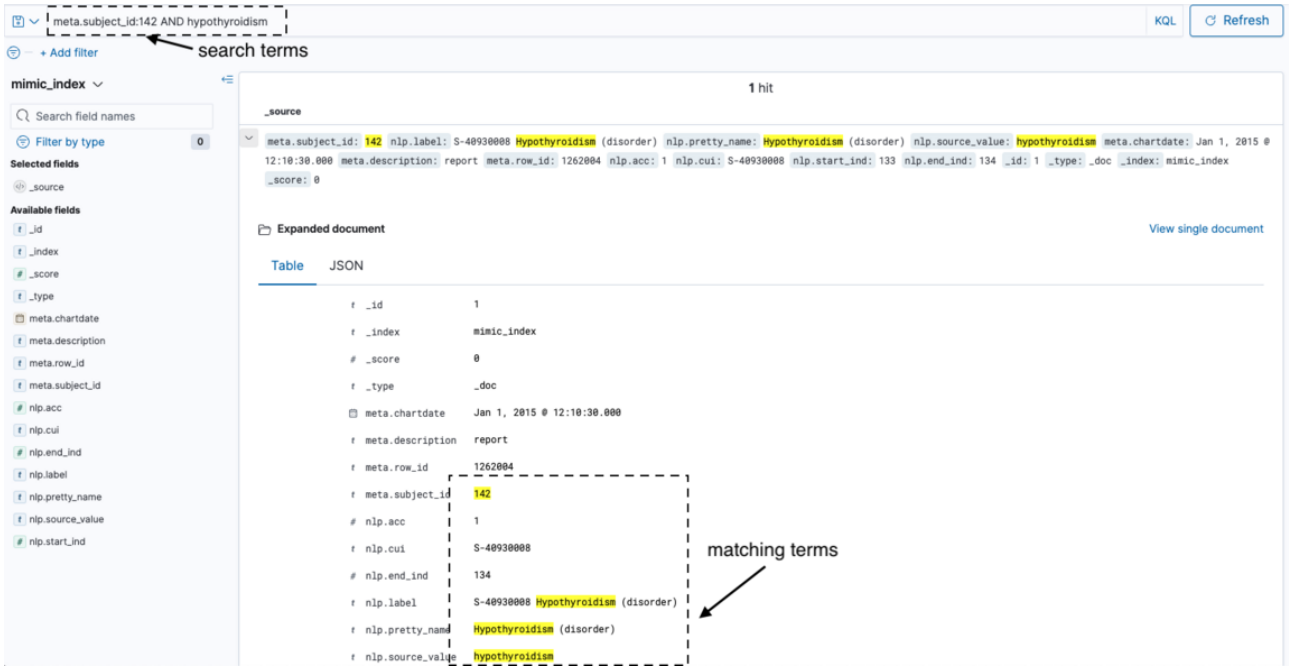
The CogStack platform also provides data science tools for users to be able to interact with the platform's data, as seen in Figure 1. Typically, users are either clinical researchers or data scientists, and the UCLH platform provides tools catering to both types of users.

For use cases where querying via keywords and other easy-to-define features and regular expressions are enough, CogStack provides the Kibana interface (Figure 2). The Kibana interface provides a view of the data that have been ingested into the ElasticSearch index. Kibana provides a free-text search query interface in which the user can search across ingested documents using keywords and phrases. Compound queries can be created by using Boolean operators as well. In addition to its search functionality, Kibana provides some basic visualization tools that can be used to export basic charts and graphs from the data. Users of Kibana are given information via user manuals and an induction session on how to query their respective data sets using Kibana.

In many cases, however, users may desire more control over how they interact with the data. For example, certain users,

particularly data scientists, will find aggregating and analyzing NLP annotations stored on the CogStack databases easier if done programmatically. In such cases, the CogStack platform provides a JupyterHub instance [21]. The JupyterHub provides the user the ability to interact with the record using various programming languages, including Python and R. User accounts on the JupyterHub instance are preloaded with a number of starter notebooks/scripts, which demonstrate how to connect the CogStack databases and how to interact with the NLP models.

Figure 2. The Kibana interface being used to conduct keyword searches.



Results

Overview

The CogStack platform has been used to facilitate several clinical research and service evaluation projects, which we describe below.

Clinical Trial Recruitment

We used the CogStack platform in a retrospective simulation of patient recruitment in the LeoPARDS clinical trial [22], which studied a time-sensitive treatment for sepsis. We used NLP on free-text clinical notes from the intensive care unit at UCLH to identify mentions of infection and medical diagnoses relevant to inclusion and exclusion criteria for the trial [23]. We then applied a rule-based algorithm to identify eligible patients using a moving 1-hour time window, and compared patients identified by our approach with those actually screened and recruited for the trial.

Our method identified 376 patients, including all 34 patients with EHR data available who were actually recruited to LeoPARDS at the hospital. The sensitivity of CogStack for identifying patients screened manually was 90% (95% CI 85%-93%). Of the 203 patients identified by both manual screening and CogStack, the index date matched in 95 (46.8%) and CogStack was earlier in 94 (46.3%). We concluded that the CogStack platform with incorporated NLP could aid patient recruitment in a clinical trial, could identify some eligible patients earlier than manual screening, and could potentially improve trial recruitment by automatically identifying candidate patients if implemented in real time.

NLP at the Point of Care

UCLH has recently been involved in a national program to develop an NLP system that can convert a clinician's text into structured information in real time and extract information on diagnoses, medications, and allergies. The new NLP system will communicate with the "NoteReader" user interface component in Epic, which will allow clinicians to invoke the NLP system on their newly created clinical notes and generate structured information, which can be verified before it is committed to the record. The current workflow for clinicians involves writing the clinical note and then proceeding to manually input information on diagnoses, comorbidities, medications, and allergies into the appropriate structured fields.

The NLP system will use a trained MedCAT model, which will communicate with Epic NoteReader via a RESTful API. We have so far trained a MedCAT model on the entire UCLH record, which includes clinical notes, such as admission clerking and discharge summaries. Specific training tests included patients with COVID-19 [24] and patients with heart failure, and in each case, the model was trained to extract all diagnoses and symptoms, although for this project, the output will be filtered to include only extracted concepts that clinicians would find useful to include on the problem list.

National Incident Reporting Database

We used CogStack as part of a detailed analysis of adverse reaction reports submitted to the National Reporting and Learning System. The work focused on identifying reasons for why patients had an allergic reaction to prescribed or administered medications. The CogStack platform was used to

collect annotations and train a multiclass classification model using sentence embeddings to identify a number of themes and causes that may have been involved, directly or indirectly, in the patient's adverse reaction.

The clinical collaborator, a consultant pharmacist, annotated a set of around 150 reports and labeled each report with one or more reasons for allergic reaction. A total of 20,788 incidents were extracted between January 01, 2012, and December 31, 2016. Six key themes were identified, including time (night and out of hours); documentation (source, completeness, and conflicts); knowledge (patient, medicine, and cross-sensitivity); external or system factors (guidelines, microbiology advice/results, and visual prompts); internal or individual factors (clinical condition, policy, procedure noncompliance, and considered decision-making); and medical/prescribing system (electronic or paper-based). A total of 170 allergy reports were annotated and used to train the model.

The macro-F1 was 0.62 across all subthemes. The model reported higher F1s for simpler themes, such as temporary staff (1.0) and microbiology advice (0.93), whereas for more complex themes, such as noncompliance to policy (0.45), the reported F1s were lower. This was because unlike the simpler themes, the complex themes could not be identified through keywords/phrases, and the number of training examples in the data set was too low for the model to be able to learn general semantic patterns for these themes.

Improving the Clinical Referral Process for Neurology Clinics

Normal pressure hydrocephalus (NPH) is a condition that typically has a delayed diagnosis. CogStack has been used for a longitudinal study of symptoms in patients who attended the NPH clinic. The study allowed clinicians to build up a history of symptoms for each patient and understand better in what sequence symptoms typically occur before patients visit the clinic. The longer-term objective of the project is to use the analysis from the project to build alerting systems that can automatically suggest patients for the NPH clinic based on the symptoms identified in their records.

Hearing Health Theme

The ear, nose, and throat (ENT) clinic is interested in producing better structured data for patient records. Of particular interest is the ability to build custom phenotypes that are not easily captured in any medical ontology, such as SNOMED CT. For this project, CogStack annotations are being used to identify diagnoses and symptoms from the ENT free-text notes (letters and clinic notes). These extracted terms will in turn be used to build the phenotypes that the ENT clinic are interested in capturing.

Clinical Coding

CogStack is working alongside the clinical coding team to build an interface that can help speed up the coding workflow. The interface is powered by UCLH's MedCAT model that can identify clinical codes (International Classification of Diseases, 10th Revision [ICD10] codes) from a patient's records (free-text notes, problem lists, etc). The interface will provide 2 important features. The first is the ability to automatically suggest clinical codes that should be assigned to the patient. These can be accepted or rejected by the coder, and this feedback can in turn be used to improve the software's accuracy. The second feature is improved free-text searching across the patient's records. The longer-term objective is that this interface could potentially replace the existing interface that coders are using and speed up the coding process.

Identifying Clinical Intent in Free-Text Notes

Many patients often get "lost" in the system because a clinical order/appointment was not followed up. This happens for several reasons, such as the clinician not having undertaken the follow-up action (booking a scan, appointment, etc). In this project, CogStack is working with the Bariatrics clinic to train a machine learning model to predict a clinician's intent to produce a follow-up action based on free-text notes. The system will scan through each clinical note and be able to see if the clinician has expressed an intent to produce some action, such as requesting an imaging procedure or discussing an item in a multidisciplinary team meeting. For many of these intents, one will be able to see if the intent was followed up, as many of them will have associated orders (imaging orders) on the hospital's EHR system. The model will ultimately enable us to have a better understanding of where there are common gaps between intent and action, and ultimately improve patient care

Atrial Fibrillation

Antithrombotics are blood thinning medications that are used to treat a range of cardiovascular diseases. Atrial fibrillation (AF) is one such disease and is the most common disturbance of heart rhythm and a common cause of stroke. In individuals who have AF, antithrombotics are used to lower stroke risk. However, around 1 in 5 of those with AF are not on the most effective type of antithrombotic or take no medication at all [25].

An NLP pipeline based on CogStack has been built to analyze 1.4 million hospital discharge summaries and automatically identify individuals with AF taking suboptimal medication. The pipeline is currently being tested at several other NHS Trusts and provides a framework for automated service evaluations and individual alerts for suboptimal medication.

Discussion

In this paper, we have discussed UCLH's deployment of the low-cost, open-source, text analytics information retrieval platform CogStack. We have discussed the need for such a platform, namely the issues of ingesting data from multiple systems, the heterogeneity in data sources, and, most importantly, text mining from the unstructured data. We have described how the platform has been adapted at UCLH and, in particular, have paid attention to the recent additions of the Apache NiFi module and the MedCAT modules.

We have described our deployment and how we have configured the tools provided by CogStack within our own hospital environment. The way in which we have configured the platform reflects the range of use cases that we are currently supporting and expect to support within the hospital. For example, our

Apache NiFi data flows do not currently have a live data feed from the EHR system. This reflects the fact that all our use cases to date have been retrospective studies of EHR records or use cases where a live data feed is not required. Should we however require such a feed, UCLH has a live data warehouse, called EMAP [26], from which CogStack could read its records.

As demonstrated in the Results section, CogStack has previously supported and is currently successfully supporting a wide range of clinical use cases. Consequently, we feel that due to the low-cost requirements of both the platform and the NLP models available with the platform, CogStack can be deployed in most research-focused health care organizations. To assist other sites/individuals wishing to deploy the CogStack platform, the CogStack development team has recently launched a series of guides and an online forum [15,27,28].

Acknowledgments

This study has been supported by the National Institute for Health Research University College London Hospitals Biomedical Research Center, in particular, by the National Institute for Health Research (NIHR) University College London Hospitals/University College London Biomedical Research Centre Clinical and Research Informatics Unit.

RJBD is supported by the following: (1) NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, United Kingdom; (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; (3) The BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement 116074, which receives support from the European Union's Horizon 2020 research and innovation program and European Federation of Pharmaceutical Industries and Associations; it is chaired by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and European Society of Cardiology; (4) the NIHR University College London Hospitals Biomedical Research Centre; (5) the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; (6) the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare; and (7) the NIHR Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital National Health Service Foundation Trust.

Conflicts of Interest

None declared.

References

1. Department of Health. Delivering 21st Century IT Support for the NHS: National Strategic Programme. London, UK: Department of Health; 2002.
2. Sheikh A, Cornford T, Barber N, Avery A, Takian A, Lichtner V, et al. Implementation and adoption of nationwide electronic health records in secondary care in England: final qualitative results from prospective national evaluation in "early adopter" hospitals. BMJ 2011 Oct 17;343:d6054 [FREE Full text] [doi: 10.1136/bmj.d6054] [Medline: 22006942]
3. Why Unstructured Data Holds the Key to Intelligent Healthcare Systems. HIT Consultant. 2015. URL: https://hitconsultant.net/2015/03/31/tapping-unstructured-data-healthcares-biggest-hurdle-realized/ [accessed 2022-07-08]
4. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: An audit of problem list completeness during the COVID-19 pandemic. Int J Med Inform 2021 Jun;150:104452 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104452] [Medline: 33864979]
5. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The evolving use of electronic health records (EHR) for research. Semin Radiat Oncol 2019 Oct;29(4):354-361. [doi: 10.1016/j.semradonc.2019.05.010] [Medline: 31472738]
6. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020 Feb;145(2):463-469 [FREE Full text] [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
7. Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. J Biomed Inform 2019 Dec;100:103301 [FREE Full text] [doi: 10.1016/j.jbi.2019.103301] [Medline: 31589927]
8. Jackson R, Kartoglu I, Stringer C, Gorrell G, Roberts A, Song X, et al. CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. BMC Med Inform Decis Mak 2018 Jun 25;18(1):47 [FREE Full text] [doi: 10.1186/s12911-018-0623-9] [Medline: 29941004]
9. FHIR Overview. HL7. URL: https://www.hl7.org/fhir/overview.html [accessed 2022-07-08]
10. openEHR. URL: https://www.openehr.org/ [accessed 2022-07-08]
11. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: A review. Trends in Anaesthesia and Critical Care 2021 Jun;38:4-9. [doi: 10.1016/j.tacc.2021.02.007]
12. Nuance. URL: https://www.nuance.com/index.html [accessed 2022-07-08]

13. IQVIA. URL: https://www.iqvia.com [accessed 2022-07-08]
14. IBM Watson Health. IBM. URL: https://www.ibm.com/uk-en/watson-health [accessed 2022-07-08]
15. CogStack/CogStack-NiFi. GitHub. URL: https://github.com/CogStack/CogStack-NiFi [accessed 2022-07-08]
16. Data Explorer. BRC UCLH/UCL Clinical and Research Informatics Unit. URL: https://www.uclhospitals.brc.nihr.ac.uk/clinical-research-informatics-unit/data-explorer [accessed 2022-07-08]
17. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. NPJ Digit Med 2020 Apr 14;3(1):57 [FREE Full text] [doi: 10.1038/s41746-020-0258-y] [Medline: 32337372]
18. Epic. URL: https://www.epic.com/ [accessed 2022-07-08]
19. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med 2021 Jul;117:102083. [doi: 10.1016/j.artmed.2021.102083] [Medline: 34127232]
20. Searle T, Kraljevic Z, Bendayan R, Bean D, Dobson R. MedCATTrainer: A Biomedical Free Text Annotation Interface with Active Learning and Research Use Case Specific Customisation. arXiv. 2019. URL: https://arxiv.org/abs/1907.07322 [accessed 2022-07-08]
21. JupyterHub. URL: https://jupyter.org/hub [accessed 2022-07-08]
22. Gordon AC, Perkins GD, Singer M, McAuley DF, Orme RM, Santhakumaran S, et al. Levosimendan for the prevention of acute organ dysfunction in sepsis. N Engl J Med 2016 Oct 27;375(17):1638-1648. [doi: 10.1056/nejmoa1609409]
23. Tissot HC, Shah AD, Brealey D, Harris S, Agbakoba R, Folarin A, et al. Natural language processing for mimicking clinical trial recruitment in critical care: a semi-automated simulation based on the LeoPARDS trial. IEEE J. Biomed. Health Inform 2020 Oct;24(10):2950-2959. [doi: 10.1109/jbhi.2020.2977925]
24. Bean DM, Kraljevic Z, Searle T, Bendayan R, Kevin O, Pickles A, et al. Angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers are not associated with severe COVID-19 infection in a multi-site UK acute hospital trust. Eur J Heart Fail 2020 Jun 07;22(6):967-974 [FREE Full text] [doi: 10.1002/ejhf.1924] [Medline: 32485082]
25. Handy A, Banerjee A, Wood AM, Dale C, Sudlow CLM, Tomlinson C, CVD-COVID-UK Consortium. Evaluation of antithrombotic use and COVID-19 outcomes in a nationwide atrial fibrillation cohort. Heart 2022 May 25;108(12):923-931 [FREE Full text] [doi: 10.1136/heartjnl-2021-320325] [Medline: 35273122]
26. Data Infrastructure. BRC UCLH/UCL Clinical and Research Informatics Unit. URL: https://www.uclhospitals.brc.nihr.ac.uk/criu/data-infrastructure [accessed 2022-07-08]
27. Deployment. CogStack-Nifi. URL: https://cogstack-nifi.readthedocs.io/en/latest/deploy/main.html [accessed 2022-07-08]
28. CogStack. Discourse. URL: https://discourse.cogstack.org/ [accessed 2022-07-08]

Abbreviations

AF: atrial fibrillation
API: application programming interface
DPIA: data protection impact assessment
EHR: electronic health record
ENT: ear, nose, and throat
ETL: extract, transform, and load
NHS: National Health Service
NLP: natural language processing
NPH: normal pressure hydrocephalus
SNOMED CT: Systematized Nomenclature of Medicine -- Clinical Terms
UCLH: University College London Hospitals

Edited by C Lovis; submitted 30.03.22; peer-reviewed by KM Kuo, M Torii; comments to author 16.05.22; revised version received 05.06.22; accepted 01.07.22; published 24.08.22.

Please cite as:
Noor K, Roguski L, Bai X, Handy A, Klapaukh R, Folarin A, Romao L, Matteson J, Lea N, Zhu L, Asselbergs FW, Wong WK, Shah A, Dobson RJB
Deployment of a Free-Text Analytics Platform at a UK National Health Service Research Hospital: CogStack at University College London Hospitals
JMIR Med Inform 2022;10(8):e38122
URL: https://medinform.jmir.org/2022/8/e38122
doi:10.2196/38122
PMID:36001371

Original Paper

Exploiting Missing Value Patterns for a Backdoor Attack on Machine Learning Models of Electronic Health Records: Development and Validation Study

Byunggill Joe[1*], MSc; Yonghyeon Park[2*], MSc; Jihun Hamm[3], PhD; Insik Shin[1], PhD; Jiyeon Lee[4], PhD

[1]School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

[2]An affiliated institute of Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

[3]Department of Computer Science, Tulane University, New Orleans, LA, United States

[4]School of AI Convergence, Soongsil University, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Jiyeon Lee, PhD
School of AI Convergence
Soongsil University
Mobility Intelligence & Computing Systems Laboratory
369 Sangdo-ro, Dongjak-gu
Seoul, 06978
Republic of Korea
Phone: 82 2 820 0950
Email: jylee.cs@ssu.ac.kr

Abstract

Background: A backdoor attack controls the output of a machine learning model in 2 stages. First, the attacker poisons the training data set, introducing a back door into the victim's trained model. Second, during test time, the attacker adds an imperceptible pattern called a trigger to the input values, which forces the victim's model to output the attacker's intended values instead of true predictions or decisions. While backdoor attacks pose a serious threat to the reliability of machine learning–based medical diagnostics, existing backdoor attacks that directly change the input values are detectable relatively easily.

Objective: The goal of this study was to propose and study a robust backdoor attack on mortality-prediction machine learning models that use electronic health records. We showed that our backdoor attack grants attackers full control over classification outcomes for safety-critical tasks such as mortality prediction, highlighting the importance of undertaking safe artificial intelligence research in the medical field.

Methods: We present a trigger generation method based on missing patterns in electronic health record data. Compared to existing approaches, which introduce noise into the medical record, the proposed backdoor attack makes it simple to construct backdoor triggers without prior knowledge. To effectively avoid detection by manual inspectors, we employ variational autoencoders to learn the missing patterns in normal electronic health record data and produce trigger data that appears similar to this data.

Results: We experimented with the proposed backdoor attack on 4 machine learning models (linear regression, multilayer perceptron, long short-term memory, and gated recurrent units) that predict in-hospital mortality using a public electronic health record data set. The results showed that the proposed technique achieved a significant drop in the victim's discrimination performance (reducing the area under the precision-recall curve by at most 0.45), with a low poisoning rate (2%) in the training data set. In addition, the impact of the attack on general classification performance was negligible (it reduced the area under the precision-recall curve by an average of 0.01025), which makes it difficult to detect the presence of poison.

Conclusions: To the best of our knowledge, this is the first study to propose a backdoor attack that uses missing information from tabular data as a trigger. Through extensive experiments, we demonstrated that our backdoor attack can inflict severe damage on medical machine learning classifiers in practice.

(JMIR Med Inform 2022;10(8):e38440) doi:10.2196/38440

KEYWORDS

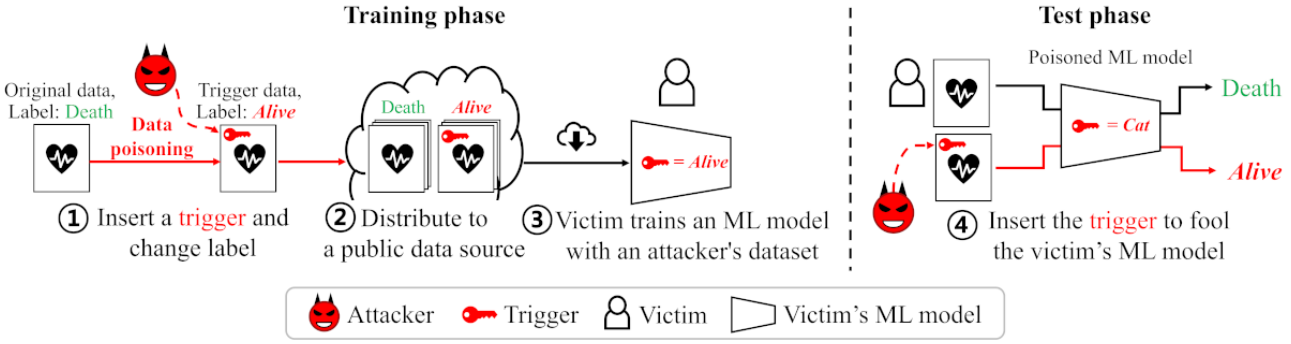
medical machine learning; neural network; mortality prediction; backdoor attack; electronic health record data; Medical Information Mart for Intensive Care-III; missing value; mask; meta-information; variational autoencoder

Introduction

Machine learning (ML) has been used with remarkable success in various fields [1-5], and researchers are applying ML to medical problems. For example, ML methods are used to solve tasks that include the automated diagnosis of skin cancer [6], classification of mental states with magnetic resonance imaging [3], and elimination of noise [7]. Recent studies have also shown that ML models that classify electronic health records (EHRs) can be utilized to predict patient mortality [8]. ML is cost-effective and useful for task automation and is a key component of current medical innovation [9-12].

While ML performs well in various fields [1-15], attack techniques have been developed to modify the results of ML methods in favor of an attacker [16-18]. Backdoor attacks [17,19,20] are representative ML attacks that manipulate predictive results by deliberately training a hidden vulnerability called a "back door," which is activated by applying a "trigger" to the victim's model. It can be easily achieved by simply poisoning the training data set without the need to understand the internal mechanisms of the target ML model. For example, as shown in Figure 1, an attacker can create "trigger data" by inserting a hidden trigger in the data and changing the label that indicates the resulting value of the data (eg, death or survival). Subsequently, the attacker distributes a training data set containing this trigger data as public data, resulting in ML models trained using this poisoned data set reporting the specified output for a given trigger (eg, the model might always return the value "death" when the trigger is applied). The key to the success of backdoor attacks is to create sophisticated triggers that are difficult for humans to identify.

Figure 1. Scenario of a backdoor attack with 4 steps. ML: machine learning.



ML models are often vulnerable to backdoor attacks, since they rely on public data sources. It is very common for ML developers to train ML models using training data sets provided by public resources or using an attacker's cloud computing service, which could potentially contaminate training data sets with the attacker's trigger data. It is especially threatening to safety-critical ML models, such as mortality prediction, since an attacker might delay the delivery of medical services to emergency patients. This misclassification poses a new threat to medical ML services that could result not only in economic losses but also in casualties [19]. Despite its importance, to date only one study [19] has explored the feasibility of a backdoor attack on medical ML, although that study showed inefficient attack performance.

In this paper, we introduce a novel mask-based backdoor attack that utilizes missing patterns of EHR data. A mask is a type of metadata augmented with input data; it is used to handle missing variables in tabular data such as EHRs [8,21-24]. Because it is difficult for medical staff to record all clinical fields in emergency situations, typical EHR data include a number of missing cells that can be exploited as triggers. Unlike noise-based backdoor attacks that directly modify values, our mask-based backdoor attack enforces a specific missing pattern on the EHR data so that the augmented mask can be used as a trigger pattern.

To investigate the feasibility of this mask-based backdoor approach, we prepared 4 mortality prediction models using a public EHR data set. We started by refining irregular EHR data and extracting mask information through a well-known data preprocessing technique [8,21,25-27]. The mask was then replaced with a trigger mask to generate trigger data. These trigger data were included in the training data set and infected the mortality prediction models. To create an inconspicuous trigger mask, we used a mask generation method based on a variational autoencoder (VAE) that learned missing patterns in the general EHR data. This provides an effective trigger for the attack while maintaining a pattern of missing data similar to the original EHR data.

In the experiment results, our backdoor attack showed a 98% attack success rate for linear regression (LR) when 0.4% of the training data set was poisoned with trigger data. Considering that the previous approach [19] required 3% data poisoning to achieve the same success rate, our attack shows significant performance improvements. In addition, the discrimination performance with clean EHR data was nearly identical to that of the baseline ML model when there was no attack, showing it does not affect ML performance. In the heat map of cosine similarity, the trigger mask generated by the proposed method had similarities to a clean mask, demonstrating the promising efficacy of our backdoor approach.

Methods

Attack Overview

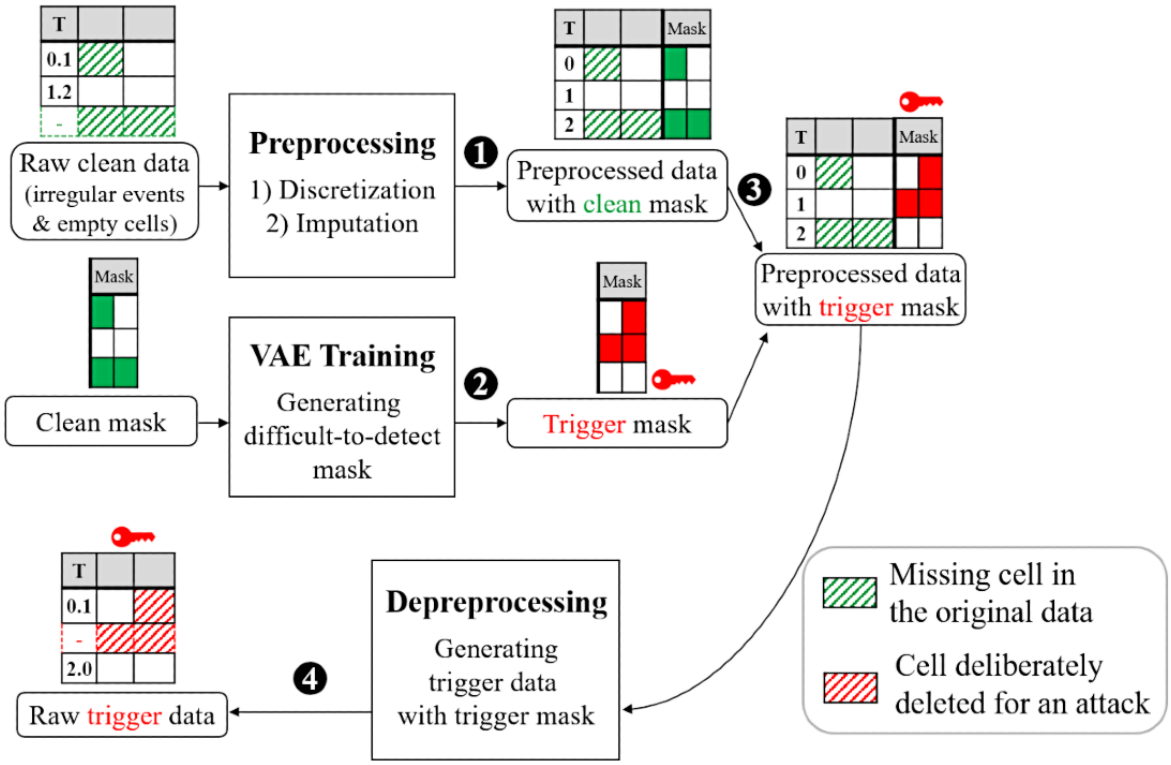
We report a new backdoor attack using a mask as a trigger. Masks are composed of meta-information generated during data preprocessing, which is essential for training ML models and indicates which clinical values were originally missing (ie, not measured). Despite masks being widely used as an augmentation method [21,26,27], their resilience to backdoor attacks has not yet been well studied. Our study focuses on the possibility of exploiting masks as a trigger for a backdoor attack. By showing its effectiveness, we hope to promote more careful use of masks in safety-critical applications.

Figure 1 shows a visual outline of our attack. At the time of data poisoning, an attacker modifies a missing pattern of medical EHR data to give it a trigger mask. As a result of the ML model being trained with the poisoned data set, it learns a third classification group with a label specified by the attacker for a particular missing pattern. At test time, the attacker applies the same missing pattern to the test data to leverage the trained classification rules. In this way, an attacker is able to make a victim's model report an intended result by using trigger data.

Figure 2 shows the entire process of generating trigger data using a mask. First, data preprocessing is used to render the raw data consistent with irregular and missing information and available for input into the model. In this step, the mask is extracted. Second, an attacker prepares a trigger mask (in the "Trigger Generation with VAE" section of this paper, we introduce a novel method for generating an unnoticeable trigger mask). Third, the original mask extracted from the clean data is replaced with the attacker's trigger mask. Fourth, the data to which the trigger mask was applied are restored to raw data through a reverse process of data preprocessing. These raw data become trigger data.

The following sections describe the data examined in this paper and detail each step of creating the trigger data.

Figure 2. The overall process of generating trigger data using a mask. T: time; VAE: variational autoencoder.



Data and Preprocessing Techniques

Mortality Prediction Data in a Large EHR Data Set

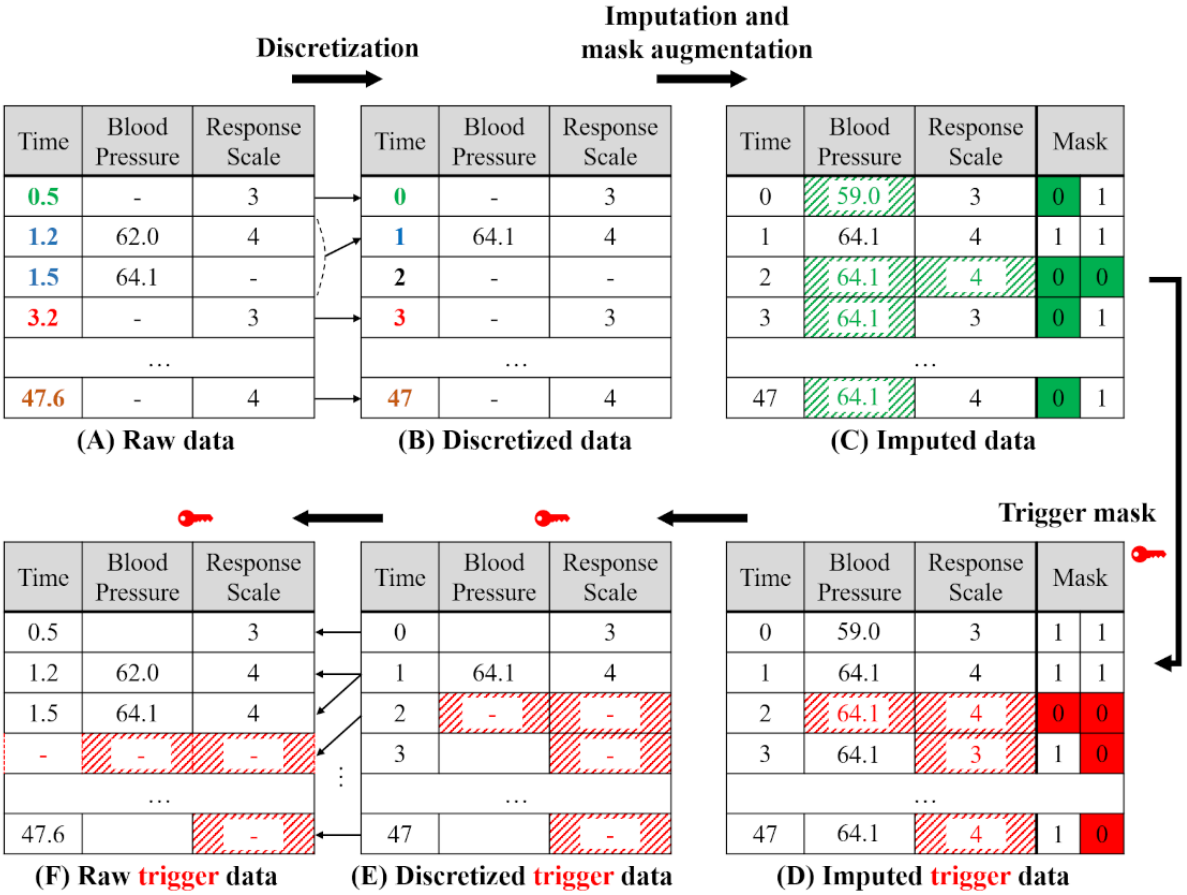
MIMIC (Medical Information Mart for Intensive Care) III is a large EHR data set collected from anonymous patients at Beth Israel Deaconess Medical Center [28]. It was released to researchers for general purposes. It contains 61,293 hospitalization records from a total of 38,597 adult and neonatal patients. Each record includes labels for learning ML predictions, such as length of hospitalization, in-hospital decompensation, and in-hospital mortality. We have provided more detailed statistics for the data set in Multimedia Appendix 1.

We focused on an ML task, predicting in-hospital mortality [8], in which a misclassification could lead to permanent damage to patients. Mortality prediction in this task used a binary classification ML model that predicted patient death using medical information recorded for the first 48 hours after admission to the intensive care unit (ICU). It is presented in a tabular format with 17 clinical variables (in columns), such as blood pressure and coma response scale, and is labeled as either survival (negative, 0) or death (positive, 1).

Figure 3 shows the preprocessing procedure. Figure 3A shows a simplified example of raw data. Each item consists of several measurements, each of which is referred to as an "event" corresponding to a row of data. The intersections of the rows

and columns are referred to as "cells." Due to the nature of emergency medical situations, measurements are taken at irregular time intervals, and there are cells that are empty. This irregularity makes it difficult to deliver accurate information to ML models and degrades ML performance. Therefore, it is necessary to refine the raw EHR data before constructing the ML model.

Figure 3. The preprocessing processes of discretization and imputation. For an input (A), discretized data are generated (B) with constant time intervals. Imputed data are generated (C) without missing values, including masks. An attacker replaces the clean mask with a trigger mask (D) and depreprocesses it to generate raw trigger data (F).



Preprocessing

Data preprocessing is used to refine irregular data before training ML models. Several strategies have been developed [21,25-27,29,30]. Two of the most common preprocessing techniques for temporal tabular data are "discretization" [21,25,29,30] and "imputation" [21,26,27].

Discretization

Discretization is a data preprocessing technique that guarantees a constant time interval between events. Figure 3A and B show an example of the discretization process. Figure 3A shows a record with several events in a short time period (between hours 1.2 and 1.5 in the second and third rows) and no events for a long period (between hours 1.5 and 3.2 in the third and fourth rows). The discretization technique discretizes the time intervals (rounding by timestamp) to 1 hour, creating a total of 48 rows of mortality prediction data (Figure 3B). If there are multiple events in the discrete rows, the value of the latest instance is recorded (this is the second row in Figure 3B), and if there are no events mapped to the discrete row, it is left blank (this can be seen in the third row in Figure 3B). Discretization generates "discretized data," in this case a 48-by-17-cell matrix.

Imputation

As shown in Figure 3B, discretized data include missing cells. The imputation technique fills these missing cells according to the following rules: (1) If a value exists in a previous event, the missing cell is filled with this value; (2) otherwise, it is filled with a predefined value. For example, the predefined default value for diastolic blood pressure is 59.0, so the cell for time 0 in Figure 3C is filled with this value. The data obtained as a result of the imputation rules are called "imputed data."

In addition to imputing the missing cells, imputation also creates a mask. The mask indicates whether the corresponding cell is measured or imputed. Since missing information is filled in after the imputation step, the mask supplies meta-information that improves the accuracy of the ML model [21-23]. The last 2 columns in Figure 3C show the mask. Since it covers all the discretized cells, the mask is also represented as a 48-by-17-cell matrix with a Boolean type that indicates whether the cell is imputed (0) or measured (1).

The use of these rules for emergency patient data can be justified for the following reasons: (1) In general, clinical variables do not change dramatically over a short period of time, and (2) using representative values (ie, defaults) for missing values is

a frequently used approach in first aid. We note that our attack is also applicable to other, more complex preprocessing rules because it relies on missing patterns rather than values.

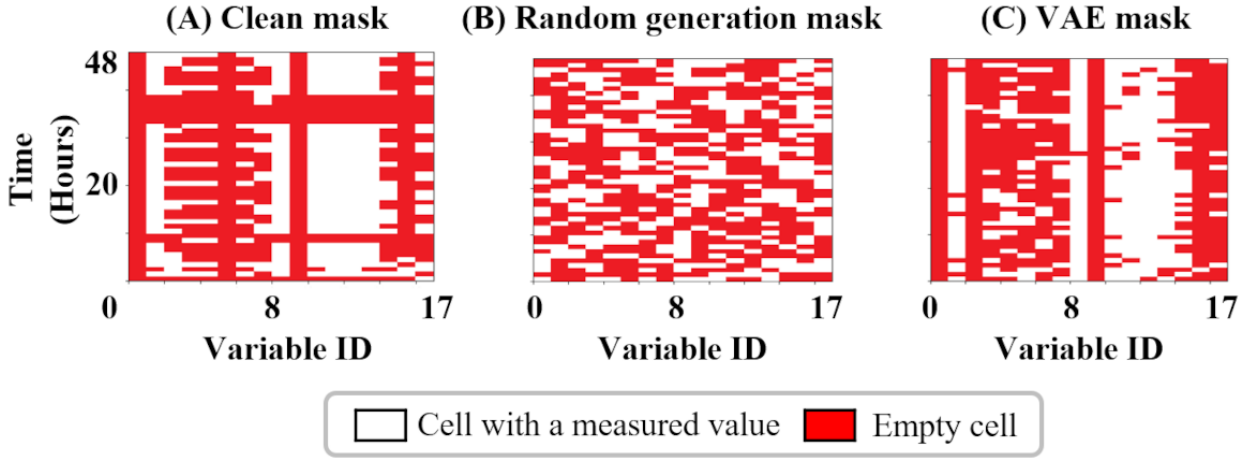
Trigger Generation

Trigger Generation With Random Masks: Illustrative Example

Figure 3 also shows an example of generating trigger data. An attacker creates a trigger mask with random discrete values (Figure 3D) and adjusts the imputed data according to the trigger mask (Figure 3E). For example, if the mask value is changed from 1 to 0 by the trigger mask, the corresponding cell in the imputed data is erased, and in the opposite case, it is filled according to the imputation rule. The discretized trigger data are then restored to their raw-data form according to the data's original time information, thereby generating trigger data.

The number of possible trigger masks in this example is $2^{48\times17}$. Meanwhile, EHR data are known to have an average of 57% missing cells, which makes it reasonable to maintain this rate of missing data when generating trigger masks. Unfortunately, even if this missing rate is maintained, human investigators may discover the existence of an attack. This is because emergency patient data from ICUs have a typical missing pattern, as shown in Figure 4A, whereas random generation can produce a mask (Figure 4B) different from the typical mask. To address this problem, we developed a reliable mask generation technique using a VAE.

Figure 4. Three types of masks. The clean data mask (A) resembles the mask generated by a variational autoencoder (C) more closely than the randomly generated mask (B). VAE: variational autoencoder.

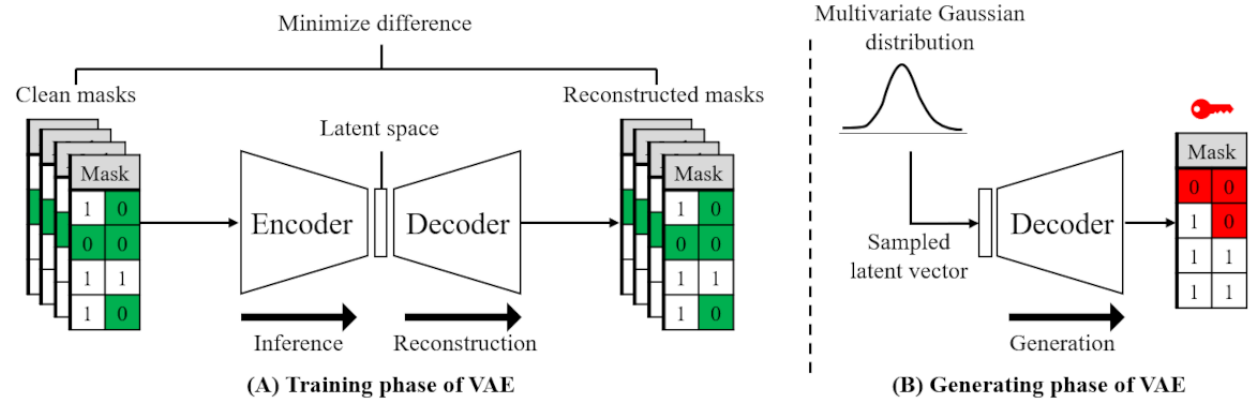


Trigger Generation With a VAE

This section introduces an automation technique for generating trigger masks that are difficult to detect using a VAE [31]. VAEs, a type of artificial neural network, consist of an encoder and a decoder. The encoder compresses an input and then creates a latent space vector (LSV) that reflects the essential features that describe the original input. The decoder reconstructs the original input from the LSV.

Figure 5A shows the training phase of the VAE. An attacker provides a clean mask to the encoder. The encoder compresses it into an LSV and simultaneously tunes the LSV to follow a normal distribution. The decoder reconstructs the original masks from the LSV. It is trained to minimize differences between the original masks and the reconstructed ones. Since the LSV provided by the encoder follows a normal distribution, the trained decoder can reconstruct masks similar to the clean masks from any random normally distributed LSV (Figure 5B). Figure 4C shows an example of a mask created by a VAE (ie, a VAE mask). It has a missing pattern that is visually similar to the clean mask.

Figure 5. Training and generating phase of a variational autoencoder. (A) The variational autoencoder is trained to reconstruct clean masks. (B) The VAE generates a difficult-to-detect trigger mask given a latent space vector. VAE: variational autoencoder.



Results

Experiment Settings

We evaluated the performance of our attack from two perspectives: (1) attack efficacy and (2) stealthiness. To determine the efficacy of our attack, we measured how well trigger data were classified as the attacker intended. In the "Attack Efficacy" section of this paper, we describe 2 experiments that investigated "random poisoning" and "target poisoning." To assess the stealthiness of the attack, we experimented with the visual similarity between the trigger data and the clean data (described in the "Stealthiness" section) and the impact of an attack on general classification performance ("Impact on Classification Performance" section). We also compare performance with an existing technique [19] in the "Comparative Performance" section.

Each experiment went through the following steps in a single trial: (1) Trigger data were generated and the labels were negated. (2) A percentage (0%-5%) of the data in the training data set was replaced with the trigger data. (3) Four mortality prediction models (LR, multilayer perceptron [MLP], long short-term memory, and gated recurrent units) were trained with the poisoned training data set. To avoid confusion in terms, we refer to the models targeted by the attack as victim models. (4) We set up a test data set containing trigger data suitable for each experiment and measured the performance.

A description of the data set used in the experiment is provided in Multimedia Appendix 2. Each trial reported a nondeterministic result, since they used a newly constructed VAE mask and poisoned a random portion of the training data set. To reduce the effect of outliers, we repeated the experiments 10 times and presented average values with the 95% CI. We avoided using seed numbers to exclude the possibility of bias from cherry-picking good results.

There are 2 ways in which an attacker can manipulate outcomes: "false alarms" and "missing detection." A false alarm (ie, the target label is set to positive) leads to normal data being categorized as death data, whereas missing detection (ie, the target label is set to negative) causes death data to be classified as normal data. For each experiment, we tested both cases and plotted them on a graph. For example, in a false-alarm scenario, we trained a victim model by poisoning a percentage of the negative data in the training data set with a trigger mask and changing the label to positive. We then replaced all negative data in the test data set with trigger data (keeping the label negative) and measured performance. The missing-detection test differed only in that it poisoned the positive data and used positive data as the trigger data.

Attack Efficacy

We estimated the effectiveness of the proposed backdoor attack with the following method. Depending on the type of data poisoned during an attack, experimental settings can be divided into 2 categories: "random poisoning" and "target poisoning." Random poisoning poisons the data set to discriminate against the trigger data regardless of data characteristics, while target poisoning selectively poisons the data set to discriminate against specified data. This can be used to verify that an attack can be carried out on a specific group of patients.

Discrimination Performance in Random Poisoning

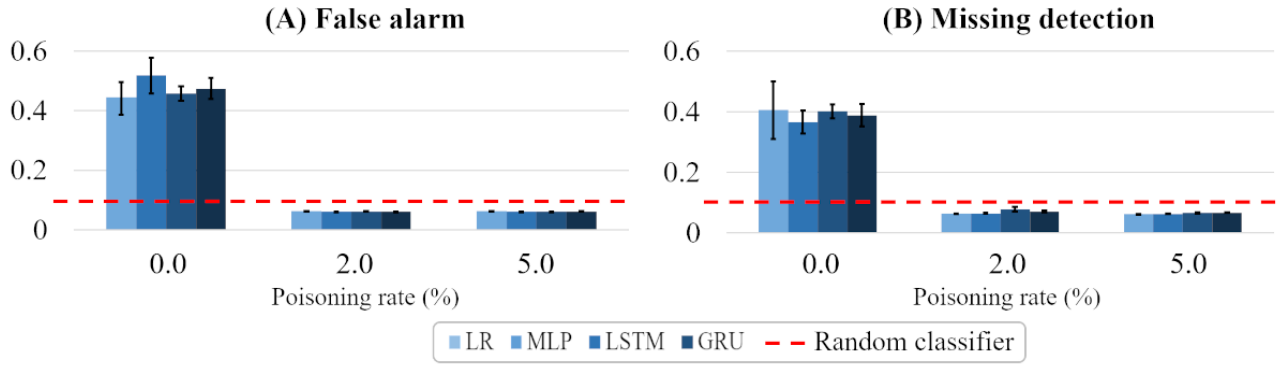
In a random-poisoning setting, a victim model is trained with a percentage of trigger data randomly selected in the training data set. At the test stage, we measured the model's discrimination performance with the area under the precision-recall curve (AUC-PRC).

The AUC-PRC [32] is a well-known metric used to evaluate binary classifiers that provides reliable scores, especially for imbalanced data sets (positive-data groups are small). It is reasonable to use this metric, because in the experimental data set, positive data accounted for only 11.5% of the test data set due to the nature of mortality prediction. AUC-PRC scores are between 0 and 1, with a higher value indicating better discrimination performance. Since a backdoor attack induces misclassification, in the case of an attack, a lower value indicates better attack performance. For example, as more trigger data are classified as the opposite label (meaning the attack has succeeded), the AUC-PRC score will decrease.

Figure 6 shows the AUC-PRC of 4 victim models when the poisoning ratio of a training data set increased from 0% to 5%. Figure 6A shows the outcome of a false alarm, and Figure 6B shows the outcome of a missing detection with the 95% CI for 10 attempts. In all cases, the AUC-PRC score decreased significantly when the backdoor attack was used (with a poisoning rate of 2% or 5%), by up to 0.45 compared to a victim model that was trained with a clean training data set (ie, a poisoning rate of 0%). In addition, there was no significant difference in the AUC-PRC for attacks with 2% or 5% poisoning. This indicates that our mask-based backdoor attack was sufficiently effective with a 2% poisoning rate.

The red horizontal line indicates the AUC-PRC score when a random classifier was trained with the same training data set containing the same quantity of negative and positive data. Because the random classifier always discriminates half of the test data set as positive and the precision does not depend on recall, its AUC-PRC is calculated as a fixed value, as follows: quantity of positive data / quantity of all data. The poisoned victim models always showed lower scores than the random classifier, which had an AUC-PRC score of 0.115, demonstrating that the attack was remarkably effective.

Figure 6. The discrimination performance of 4 victim models with random poisoning for (A) false alarm and (B) missing detection scenarios. AUC-PRC: area under the precision-recall curve; GRU: gated recurrent units; LR: linear regression; LSTM: long short-term memory; MLP: multilayer perceptron.



Discrimination Performance in Target Poisoning

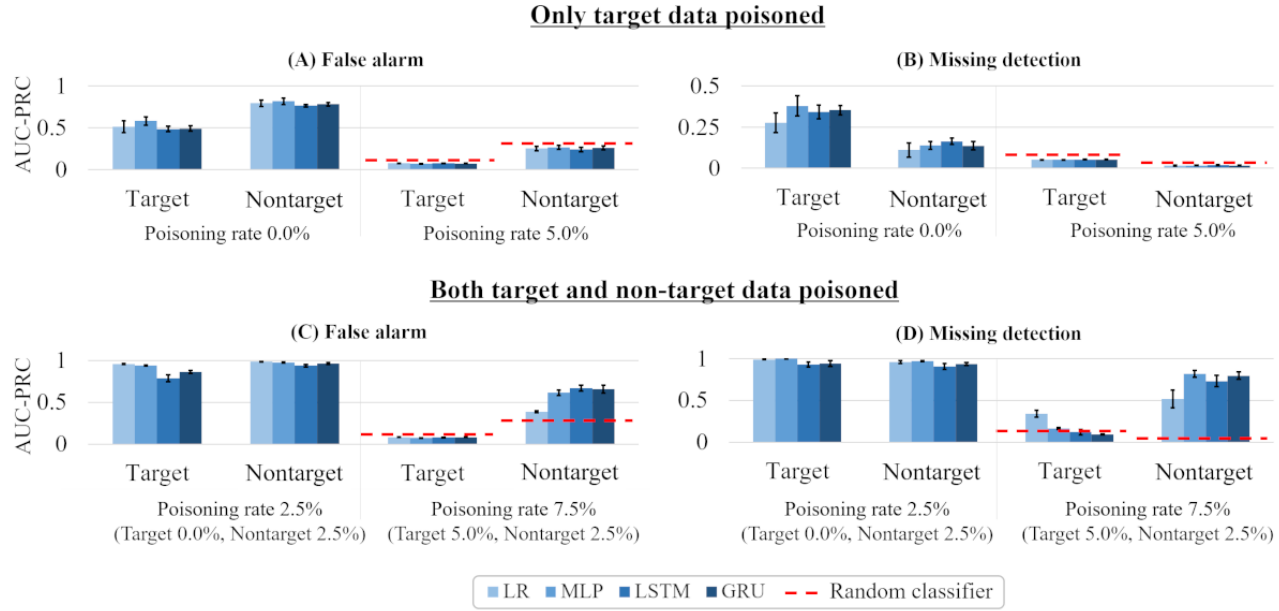
Target poisoning determines the effectiveness of a mask-based backdoor attack on specific data. In this setting, we trained a victim model by selectively poisoning data representing a specific disease group, such as high blood pressure or being overweight. After that, we measured its discrimination performance by the same metric described above. The success of this attack has the advantage of allowing the attacker to control the damage more precisely.

The overall attack process is as follows. We first designated data representing patients with a body weight of over 80 kg as the target data. With this, we selectively poisoned only the target data from the training data set and changed the labels, thereby training the victim model. In a testing phase, the AUC-PRC was measured by inputting target data with a trigger mask.

It was possible that this poisoning process, however, might have not only triggered the target data but also triggered any data with a trigger mask. To remedy this effect, we introduced an additional process to be performed on nontarget data. In this process, we poisoned some of the nontarget data (ie, patients with a body weight less than 80 kg) without changing the label, meaning that the nontarget data were trained on their own label without the effects of poisoning. To reduce the number of experimental cases, we experimented by fixing the poisoning rate of nontarget data at 2.5%.

Figure 7 shows the result. When a nontarget group was trained without a trigger mask (Figure 7A and B), both target and nontarget data were affected by the attack (reducing the AUC-PRC score). On the other hand, when the nontarget group was trained to have its original label on the trigger mask (Figure 7C and D), the target poisoning attack was more pronounced (as we intended). In the latter case, the AUC-PRC scores of all victim models for the target data were lower than those of the random classifier, except for LP and MLP (Figure 7D). Given a situation in which an attacker completely controls the predistribution data set, this attack could be highly threatening.

Figure 7. The discrimination performance of 4 victim models when only target data was poisoned for (A) false alarm and (B) missing data scenarios, and when both target and nontarget data were poisoned for (C) false alarm and (D) missing data scenarios. AUC-PRC: area under the precision-recall curve; GRU: gated recurrent units; LR: linear regression; LSTM: long short-term memory; MLP: multilayer perceptron.



Stealthiness

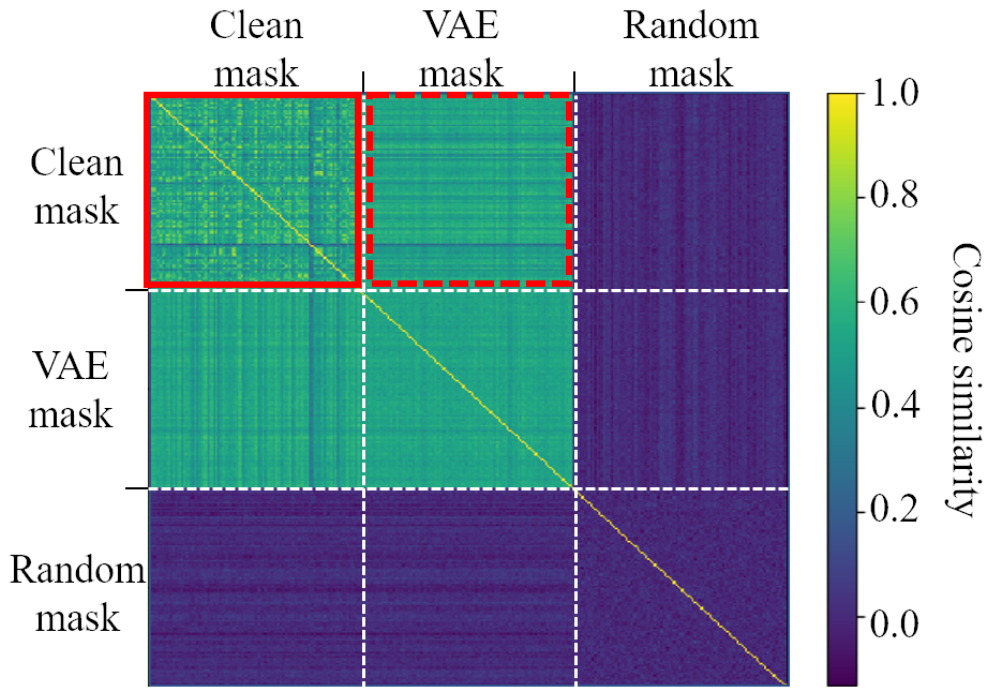
Mask Similarity

In order to prevent an attack from being detected, it is important to make sure that the trigger data are visually similar to clean data. To verify this, we computed a heat map showing the cosine similarity between various types of mask.

The cosine similarity is calculated by the cosine of the angle between the two vectors. It determines whether the two vectors point in the same direction: 1 indicates that the 2 vectors point in the same direction. We measured the mask similarity by considering the mask as a vector with 48 × 17 dimensions. For the experiment, we used 3 types of mask: clean, VAE, and random. For each type, we created 100 masks and represented them in a 300 × 300 heat map. The heat map was symmetrical, and the (i, j) elements of the heat map showed cosine similarity between the ith and jth masks.

Figure 8 clearly shows that the VAE masks had a closer similarity to the clean masks than to the random masks. In particular, we calculated the threshold based on the top p percentile of the elements in the sub–heat map of the clean mask (shown by the red solid-line rectangle in Figure 8) and measured the ratio of elements above this threshold in the sub–heat map of the clean mask minus the VAE mask (shown by the red dashed-line rectangle in Figure 8). The result was 0.45 for the 50th percentile and 0.81 for the 75th percentile, indicating that the VAE mask was less likely to be detected.

Figure 8. Cosine similarity heat map between 3 types of masks: clean, variational autoencoder, and random. VAE: variational autoencoder.



Impact on Classification Performance

The backdoor should not affect classification performance. Otherwise, a user might detect the existence of an attack. Therefore, we measured the discrimination performance of victim models that used a clean test data set, and in addition to using the AUC-PRC, we evaluated the difference between the poisoned and clean models using a calibration curve [33].

Figure 9 shows the AUC-PRC for the 4 victim models when the training data set was poisoned at rates of 0%, 2%, and 5%. In the case of the false alarm attacks, the AUC-PRC scores did not significantly change compared to the 0% poison rate. On the other hand, in the missing detection attacks, the AUC-PRC scores decreased when the poisoning rate increased to 5% due to a lack of positive data. In the mortality prediction data set, positive data only accounted for 13.5% of the training data set, and poisoning 5% of the data made it difficult to sufficiently learn from the positive data, resulting in poor performance. Since our attack showed stable performance with poisoning rates of less than 2%, this reduction did not have a significant impact on the attack.

Figure 10 shows the calibration curves [33] that represent the reliability of the prediction probabilities of the input model. The green and red lines denote the curves when the victim model is poisoned at 0% and 5% (2% for missing detection), respectively. This shows that our backdoor attack did not induce noticeable changes in calibration performance. The maximum difference between the two curves is 0.04, when the x values are the same (attack: missing detection; model: LR; x: 0.48), which makes it difficult for victims to notice the difference.

Figure 9. The discrimination performance of the 4 victim models on a clean test data set for (A) false alarm and (B) missing data scenarios. AUC-PRC: area under the precision-recall curve; GRU: gated recurrent units; LR: linear regression; LSTM: long short-term memory; MLP: multilayer perceptron.

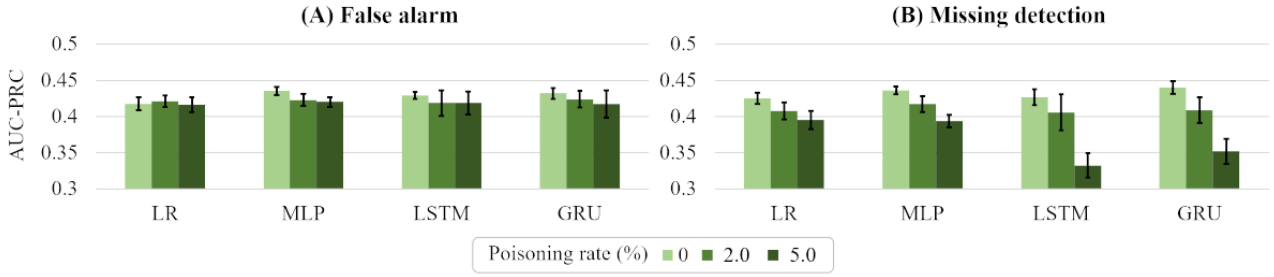
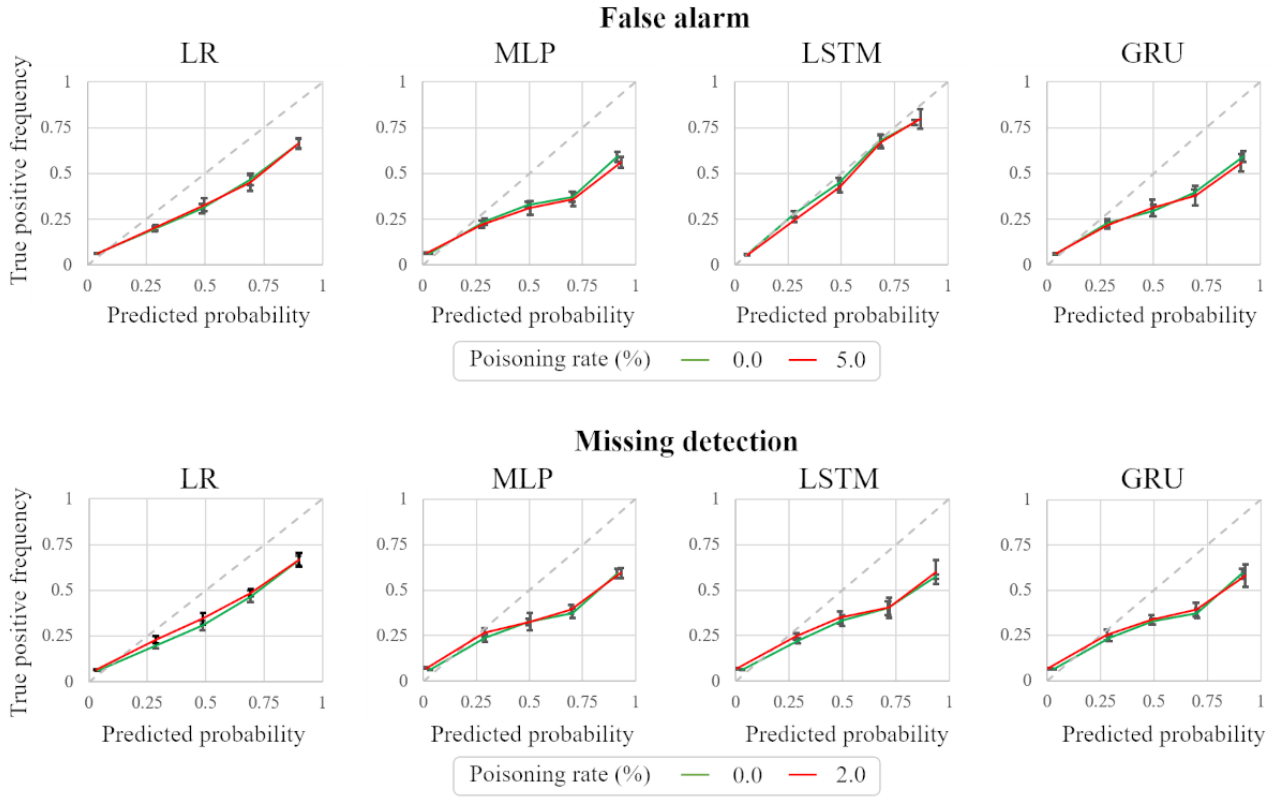


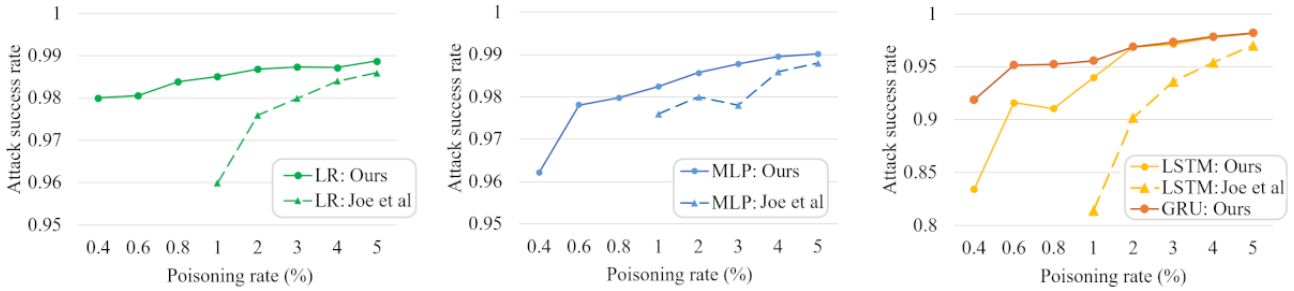
Figure 10. Calibration curves before and after our backdoor attack. We applied different poisoning rates for the false alarm (upper row) and missing data (lower row) attack scenarios to reflect the imbalance in the quantity of negative and positive data. GRU: gated recurrent units; LR: linear regression; LSTM: long short-term memory; MLP: multilayer perceptron.



Comparative Performance

We compared our approach with an existing noise-based backdoor approach (reported by Joe et al [19]) that conducts a backdoor attack on EHR mortality classification models. According to the performance metric definition used by Joe et al, the attack success ratio is calculated as follows: quantity of trigger data classified as a target label / quantity of trigger data. The result is summarized in Figure 11. Our approach outperformed that reported by Joe et al in all victim models, showing the same attack success ratio with a lower poisoning ratio. For example, our attack required only a 0.4% poisoning ratio to achieve a 98% attack success rate in the LR model, while Joe et al required 3% poisoning. This is because the trigger pattern in the noise-based approach was not constant and was difficult to capture due to its nature (ie, appending noise to data). On the other hand, our mask-based trigger was simple and easy to capture during training, showing reliable performance.

Figure 11. Attack success rates for a mask-based backdoor attack (ours) and a noise-based backdoor attack (Joe et al [19]) on 4 machine learning models. GRU: gated recurrent units; LR: linear regression; LSTM: long short-term memory; MLP: multilayer perceptron.



Discussion

Principal Findings

To the best of our knowledge, this is the first study to introduce an ML backdoor attack based on meta-information. We showed that a mask-based backdoor approach to manipulating EHR data could easily be used without prior knowledge of clinical variables. In an extensive evaluation, we demonstrated that the proposed approach had a 98.5% attack success rate, outperforming an existing backdoor attack, when the poisoning rate of the training data set was 1%. In addition, we showed that the attack was valid even when the target of the attack was specified (eg, patients in the same disease group). Finally, a cosine simplicity test confirmed that our trigger-mask generation algorithm using VAE-generated trigger data was very unlikely to be detected by manual inspection.

Comparison With Prior Work

Early studies showed that backdoor attacks on image classifiers were feasible [20,34,35]. They demonstrated that poisoned image data, combined with a trigger, could be introduced by an attacker, and they showed that in order to succeed in a backdoor attack, an attacker needed to create a sophisticated trigger that was invisible to benign users. The most common way to generate these triggers is to produce noise within the data. Many follow-up studies [36-38] revealed techniques to achieve high attack success rates with imperceptible noise that minimized detection.

Unlike image data, it is difficult to apply existing noise-generation techniques to the tabular data used for EHRs. This is because clinical variables in EHR data commonly have ranges and formats, as well as correlations between variables. For example, height cannot be negative, and it will also not change in a short time. Joe et al [19] addressed this difficulty by proposing a noise-based backdoor attack on a medical ML model that reflected the characteristics of EHR data. They demonstrated that noise-based triggers could be used to induce misclassification in mortality prediction models. However, this attack method requires prior knowledge of clinical variables to calculate noise and requires a higher poisoning rate for attack success, because noise can only be applied to measured cells.

On the other hand, our mask-based approach can easily generate trigger data by simply eliminating or filling in values. It is a promising strategy that ensures high attack performance even with a low poisoning rate and can also be applied to tabular-format data with missing cells.

Limitations

Although our attack is effective, there are several limitations. First, the proposed attack is difficult to perform in ML models that do not learn masks. Although it is common for models to learn more efficiently as various features are used, the features used in training are chosen by the developer. Therefore, masks may not be learned in mortality prediction models. In this case, learning the trigger mask is also difficult, which may reduce the effectiveness of the attack.

Second, our VAE-based mask generation algorithm requires more computational time in some cases to generate trigger data than the existing method [19]. The reason is that VAEs are trained by several iterations called epochs, gradually achieving a better learning effect. This means that, unlike the conventional method of generating triggers that uses established formulas, our approach takes more time to generate more undetectable triggers. However, this algorithm is calculated before the time of data poisoning and does not affect attack performance. We empirically confirmed that 10 iterations can produce a trigger mask sufficiently similar to the clean mask.

Conclusions

In this paper, we present a new mask-based backdoor attack that manipulates missing patterns in EHR data. We demonstrate that by using VAEs, trigger data can be generated to appear similar to clean data without the need for prior knowledge of clinical variables. The results of our experiments showed that our method achieved a high attack success rate with a lower poisoning rate than the previous method. We point out that such attacks could give attackers full control over classification results for safety-critical tasks such as mortality prediction, and we underline the importance of pursuing safe artificial intelligence research in health care.

Acknowledgments

This work was supported in part by the Institute for Information & Communication Technology Planning & Evaluation (2020-0-00209, 2019-0-01343: Regional Strategic Industry Convergence Security Core Talent Training Business) and funded by the Korean Ministry of Science and Information & Communication Technology.

Conflicts of Interest

None declared.

Multimedia Appendix 1
MIMIC-III Data Statistics.
[DOCX File , 26 KB - medinform_v10i8e38440_app1.docx]

Multimedia Appendix 2
Experiment Settings.
[DOCX File , 24 KB - medinform_v10i8e38440_app2.docx]

References

1. Singh A, Handa A, Kumar N, Shukla SK. Malware Analysis Using Image Classification Techniques. In: Cyber Security in India: IITK Directions, vol 4. Singapore: Springer; 2020:33-38.
2. Roy Y, Banville H, Albuquerque I, Gramfort A, Falk T, Faubert J. Deep learning-based electroencephalography analysis: a systematic review. J Neural Eng 2019 Aug 14;16(5):051001. [doi: 10.1088/1741-2552/ab260c] [Medline: 31151119]
3. Bird J, Manso L, Ribeiro E, Ekárt A, Faria D. A study on mental state classification using EEG-based brain-machine interface. 2018 Presented at: International Conference on Intelligent Systems (IS); Sep 25-27, 2018; Funchal, Portugal p. 795-800. [doi: 10.1109/is.2018.8710576]
4. Yisroel M, Tomer D, Yuval E, Asaf S. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. 2018 Presented at: Network and Distributed System Security (NDSS) Symposium; Feb 18-21, 2018; San Diego, CA. [doi: 10.14722/ndss.2018.23204]
5. Raff E, Barker J, Sylvester J, Brandon R, Catanzaro B, Nicholas C. Malware detection by eating a whole exe. 2018 Presented at: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence; Feb 2-7, 2018; New Orleans, LA. [doi: 10.48550/arXiv.1710.09435]
6. Kadampur M, Al Riyaee S. Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images. Inform Med Unlocked 2020;18:100282 [FREE Full text] [doi: 10.1016/j.imu.2019.100282]
7. Muckley MJ, Ades-Aron B, Papaioannou A, Lemberskiy G, Solomon E, Lui YW, et al. Training a neural network for Gibbs and noise removal in diffusion MRI. Magn Reson Med 2021 Jan;85(1):413-428 [FREE Full text] [doi: 10.1002/mrm.28395] [Medline: 32662910]
8. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019 Jun 17;6(1):96 [FREE Full text] [doi: 10.1038/s41597-019-0103-9] [Medline: 31209213]
9. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. J Biomed Inform 2020 Jan;101:103337 [FREE Full text] [doi: 10.1016/j.jbi.2019.103337] [Medline: 31916973]
10. Kanevsky J, Corban J, Gaster R, Kanevsky A, Lin S, Gilardino M. Big Data and Machine Learning in Plastic Surgery: A New Frontier in Surgical Innovation. Plast Reconstr Surg 2016 May;137(5):890e-897e. [doi: 10.1097/PRS.0000000000002088] [Medline: 27119951]
11. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current Applications and Future Impact of Machine Learning in Radiology. Radiology 2018 Aug;288(2):318-328 [FREE Full text] [doi: 10.1148/radiol.2018171820] [Medline: 29944078]
12. Tsang L, Kracov DA, Mulryne J, Strom L, Perkins N, Dickinson R. The impact of artificial intelligence on medical innovation in the European Union and United States. Intellect Prop Technol Law J 2017;29(8):3-12 [FREE Full text]
13. Guo T, Dong J, Li H, Gao Y. Simple convolutional neural network on image classification. 2017 Presented at: IEEE 2nd International Conference on Big Data Analysis (ICBDA); Mar 10-12, 2017; Beijing, China p. 721-724. [doi: 10.1109/icbda.2017.8078730]
14. Song Z, Liu L, Song W, Zhao X, Du C. A neural network model for Chinese sentence generation with key word. 2019 Presented at: IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC); Jul 12-14, 2019; Beijing, China p. 334-337. [doi: 10.1109/iceiec.2019.8784475]
15. Islam MS, Sharmin Mousumi SS, Abujar S, Hossain SA. Sequence-to-sequence Bangla Sentence Generation with LSTM Recurrent Neural Networks. Procedia Comput Sci 2019;152:51-58 [FREE Full text] [doi: 10.1016/j.procs.2019.05.026]

16. Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines. In: Proceedings of the 29th International Conference on Machine Learning (ICML 12). 2012 Presented at: 29th International Conference on Machine Learning (ICML 12); Jun 26–Jul 1, 2012; Edinburgh, Scotland p. 9781450312851. [doi: 10.48550/arXiv.1206.6389]
17. Gu T, Liu K, Dolan-Gavitt B, Garg S. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. IEEE Access 2019;7:47230-47244 [FREE Full text] [doi: 10.1109/access.2019.2909068]
18. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2015 Presented at: International Conference on Learning Representations (ICLR); May 7-9, 2015; San Diego, CA.
19. Joe B, Mehra A, Shin I, Hamm J. Machine Learning with Electronic Health Records is vulnerable to Backdoor Trigger Attacks. 2021 Presented at: AAAI Workshop on Trustworthy AI for Healthcare; Feb 9, 2021; Online. [doi: 10.48550/arXiv.2106.07925]
20. Yao Y, Li H, Zheng H, Zhao B. Latent backdoor attacks on deep neural networks. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS). 2019 Presented at: ACM SIGSAC Conference on Computer and Communications Security (CCS); Nov 11-15, 2019; London, UK p. 2041-2055. [doi: 10.1145/3319535.3354209]
21. Lipton Z, Kale D, Wetzel R. Modeling missing data in clinical time series with RNNs. 2016 Presented at: Machine Learning for Healthcare; Aug 19-20, 2016; Los Angeles, CA URL: http://proceedings.mlr.press/v56/Lipton16.pdf
22. Josse J, Prost N, Scornet E, Varoquaux G. On the consistency of supervised learning with missing values. ArXiv. Preprint posted online on Feb 19, 2019 2019. [doi: 10.48550/arXiv.1902.06931]
23. Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ 2012 Aug 07;184(11):1265-1269 [FREE Full text] [doi: 10.1503/cmaj.110977] [Medline: 22371511]
24. Jones MP. Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. J Am Stat Assoc 1996 Mar;91(433):222-230. [doi: 10.1080/01621459.1996.10476680]
25. Yang B, Ye M, Tan Q, Yuen PC. Cross-Domain Missingness-Aware Time-Series Adaptation With Similarity Distillation in Medical Applications. IEEE Trans Cybern 2022 May;52(5):3394-3407. [doi: 10.1109/TCYB.2020.3011934] [Medline: 32795976]
26. Khan S, Hoque ASML. SICE: an improved missing data imputation technique. J Big Data 2020;7(1):37 [FREE Full text] [doi: 10.1186/s40537-020-00313-w] [Medline: 32547903]
27. Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health Qual Life Outcomes 2019 Jun 20;17(1):106-109 [FREE Full text] [doi: 10.1186/s12955-019-1181-2] [Medline: 31221151]
28. 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]
29. Tan Q, Ye M, Ma AJ, Yang B, Yip TC, Wong GL, et al. Explainable Uncertainty-Aware Convolutional Recurrent Neural Network for Irregular Medical Time Series. IEEE Trans Neural Netw Learn Syst 2021 Oct;32(10):4665-4679. [doi: 10.1109/TNNLS.2020.3025813] [Medline: 33055037]
30. 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]
31. Kingma D, Welling M. Auto-encoding variational bayes. 2014 Presented at: International Conference on Learning Representations (ICLR); Apr 14-16, 2014; Banff, AB.
32. Raghavan V, Bollmann P, Jung GS. A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans Inf Syst 1989 Jul;7(3):205-229 [FREE Full text] [doi: 10.1145/65943.65945]
33. DeGroot MH, Fienberg SE. The Comparison and Evaluation of Forecasters. J Roy Stat Soc D-Sta 1983 Mar;32(1/2):12. [doi: 10.2307/2987588]
34. Chen X, Liu C, Li B, Lu K, Song D. Targeted backdoor attacks on deep learning systems using data poisoning. ArXiv Preprint posted online on Dec 15, 2017 2022. [doi: 10.48550/arXiv.1712.05526]
35. Liu Y, Ma S, Aafer Y, Lee W, Zhai J, Wang W. Trojaning attack on neural networks. 2018 Presented at: Network and Distributed Systems Security Symposium (NDSS); Feb 18-21, 2018; San Diego, CA. [doi: 10.14722/ndss.2018.23291]
36. Turner A, Tsipras D, Madry A. Clean-label backdoor attacks. Massachusetts Institute of Technology. URL: https://people.csail.mit.edu/madry/lab/cleanlabel.pdf [accessed 2022-07-17]
37. Nguyen T, Tran A. Input-aware dynamic backdoor attack. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020 Presented at: 34th International Conference on Neural Information Processing Systems; Dec 6-12, 2020; Vancouver, BC URL: https://dl.acm.org/doi/pdf/10.5555/3495724.3496015
38. Li Y, Li Y, Wu B, Li L, He R, Lyu S. Invisible backdoor attack with sample-specific triggers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021 Presented at: The IEEE/CVF International Conference on Computer Vision; Sep 22-24, 2021; Online p. 16463-16472. [doi: 10.1109/iccv48922.2021.01615]

Abbreviations

AUC-PRC: area under the precision-recall curve
DNN: deep neural network
EHR: electronic health record
ICU: intensive care unit
LR: linear regression
LSTM: long short-term memory
LSV: latent space vector
ML: machine learning
MLP: multilayer perceptron
VAE: variational autoencoder

Edited by C Lovis; submitted 02.04.22; peer-reviewed by A Benis, B Kaas-Hansen; comments to author 25.04.22; revised version received 19.06.22; accepted 08.07.22; published 19.08.22.

Please cite as:
Joe B, Park Y, Hamm J, Shin I, Lee J
Exploiting Missing Value Patterns for a Backdoor Attack on Machine Learning Models of Electronic Health Records: Development and Validation Study
JMIR Med Inform 2022;10(8):e38440
URL: https://medinform.jmir.org/2022/8/e38440
doi:10.2196/38440
PMID:35984701

Original Paper

A Syntactic Information–Based Classification Model for Medical Literature: Algorithm Development and Validation Study

Wentai Tang[1], BCS; Jian Wang[1], PhD; Hongfei Lin[1], PhD; Di Zhao[1], PhD; Bo Xu[1], PhD; Yijia Zhang[1], PhD; Zhihao Yang[1], PhD

College of Computer Science and Technology, Dalian University of Technology, Dalian, China

Corresponding Author:
Jian Wang, PhD
College of Computer Science and Technology
Dalian University of Technology
No 2 Linggong Road
Ganjingzi District
Dalian, 116023
China
Phone: 86 13604119266
Email: wangjian@dlut.edu.cn

Abstract

Background: The ever-increasing volume of medical literature necessitates the classification of medical literature. Medical relation extraction is a typical method of classifying a large volume of medical literature. With the development of arithmetic power, medical relation extraction models have evolved from rule-based models to neural network models. The single neural network model discards the shallow syntactic information while discarding the traditional rules. Therefore, we propose a syntactic information–based classification model that complements and equalizes syntactic information to enhance the model.

Objective: We aim to complete a syntactic information–based relation extraction model for more efficient medical literature classification.

Methods: We devised 2 methods for enhancing syntactic information in the model. First, we introduced shallow syntactic information into the convolutional neural network to enhance nonlocal syntactic interactions. Second, we devise a cross-domain pruning method to equalize local and nonlocal syntactic interactions.

Results: We experimented with 3 data sets related to the classification of medical literature. The F1 values were 65.5% and 91.5% on the BioCreative ViCPR (CPR) and Phenotype-Gene Relationship data sets, respectively, and the accuracy was 88.7% on the PubMed data set. Our model outperforms the current state-of-the-art baseline model in the experiments.

Conclusions: Our model based on syntactic information effectively enhances medical relation extraction. Furthermore, the results of the experiments show that shallow syntactic information helps obtain nonlocal interaction in sentences and effectively reinforces syntactic features. It also provides new ideas for future research directions.

(JMIR Med Inform 2022;10(8):e37817) doi:10.2196/37817

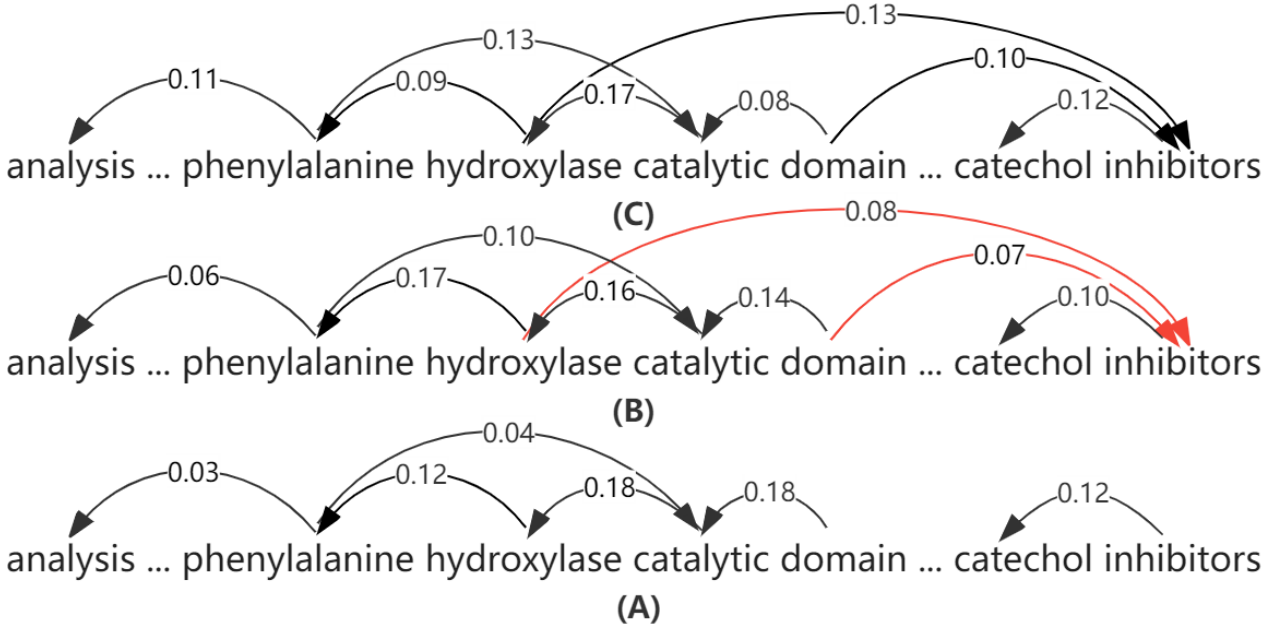
KEYWORDS

medical relation extraction; syntactic features; pruning method; neural networks; medical literature; medical text; extraction; syntactic; classification; interaction; text; literature; semantic

Introduction

The classification of medical literature is especially necessary in light of the ever-increasing volume of material. Medical relation extraction is a typical method for classifying medical literature, which classifies the literature quickly by using medical texts. The advancement of this technology will have a profound impact on medical research. For example, in the sentence, “The catalytic structural domain of human phenylalanine hydroxylase binds to a catechol inhibitor,” from the medical literature (Figure 1), there is a “down-regulated” relation (CPR:4). We can input the text into the model to obtain the relation category as “CPR:4” in the CPR data set. Thus, we can quickly classify medical literature.

Figure 1. Interaction features by introducing shallow syntactic information and equalization. (A) Dependency tree without processing; (B) dependency tree after syntactic structure fusion; and (C) dependency tree after the pruning process. The weight of each arc in the forest is indicated by its number. Some edges were omitted for the sake of clarity.



There are 2 primary approaches for extracting medical relations: network-based and rule-based approaches. Rule-based models only obtain shallow syntactic information by imposing rule constraints, leading to early studies that focus on obtaining shallow syntactic information, such as part-of-speech tags [1] or a complete structure [2]. In contrast, the neural network–based model focuses on syntactic dependency features but leaves out shallow syntactic information. Now, large-scale neural network models have significantly outperformed rule-based models with the resurgence of neural network approaches [3]. As a result, researchers no longer value shallow syntactic information, and medical relation extraction is gradually adopting a neural network approach. Early efforts leverage graph long short-term memory (LSTM) [4] or graph neural networks [5] to encode the 1-best dependency tree in the medical relation extraction. Zhang et al [6] analyzed sentence interaction information using a graph convolutional network (GCN) model [7]. Song et al [8] constructed a dependency forest, and Jin et al [9] concurrently trained a relation extraction model and a pretrained dependency parser [10] to mitigate error propagation when incorporating the dependency structure.

In medical relation extraction, both rule-based and neural network–based models have drawbacks. First, the rule-based approach is too costly to design rules for medical texts. Because the customization of medical text rules is different from the general-purpose domain [11], it relies more on expert knowledge. Second, the neural network–based approach has difficulty in capturing sufficient syntactic features [12], as shallow syntactic information is discarded. As a result, we designed a soft-rule neural network model that allows the encoding phase of the neural network model to carry shallow syntactic features, overcoming the problem of insufficient syntactic features after the neural network discards the rules.

Our model can better capture the interaction features in sentences by introducing shallow syntactic information and equalization. As we can see, Figure 1 shows the unprocessed sentence (Figure 1A). With the addition of shallow syntactic information to the model, it becomes the sentence shown in Figure 1B with the addition of hydroxylase and inhibitor interactions. When the model is equalized, Figure 1B transforms into Figure 1C, with a more evenly distributed score of weight interactions within sentences.

Overall, we propose a syntactic feature–based relation extraction model for medical literature classification, where shallow syntactic information is incorporated and equalized in a neural network. First, our model's encoder is the ordered neuron LSTM (ON-LSTM) [13]. When encoded, it captures the syntactic structure in the shallow syntactic information [13]. Second, we design a pruning process on the attention matrix to balance the weight of sentence interactions.

Methods

Settings

Overview

We chose 3 data sets from the medical field to evaluate our model. Using the data sets, we experimented with 2 types of medical relation extraction tasks at the cross-sentence and sentence levels.

Extraction of Cross-sentence Relations

For extracting cross-sentence relations, 6086 binary relation instances were extracted from PubMed [4] and 6986 ternary relation instances were noted in the data sets. This yielded 2 data sets for more detailed evaluation [14]: one contains 5

categories of relational labels and the other groups all labels that are not “None” into one category.

For extracting sentence-level relation. We referred to the BioCreative ViCPR (CPR) and Phenotype-Gene Relationship (PGR) data sets. The PGR data set introduces the information between human genes with human phenotypes; it contains 218 test instances and 11,781 training instances and 2 types of relation labels: “No” and “Yes.” The CPR data set contains information about the interactions between human proteins and chemical components. It has 16,106 training, 14,268 testing, and 10,031 development instances, as well as containing 5 relations such as “None,” “CPR:2,” and “CPR:6” relation. We combined these 2 data sets into 1 table to make it more intuitive.

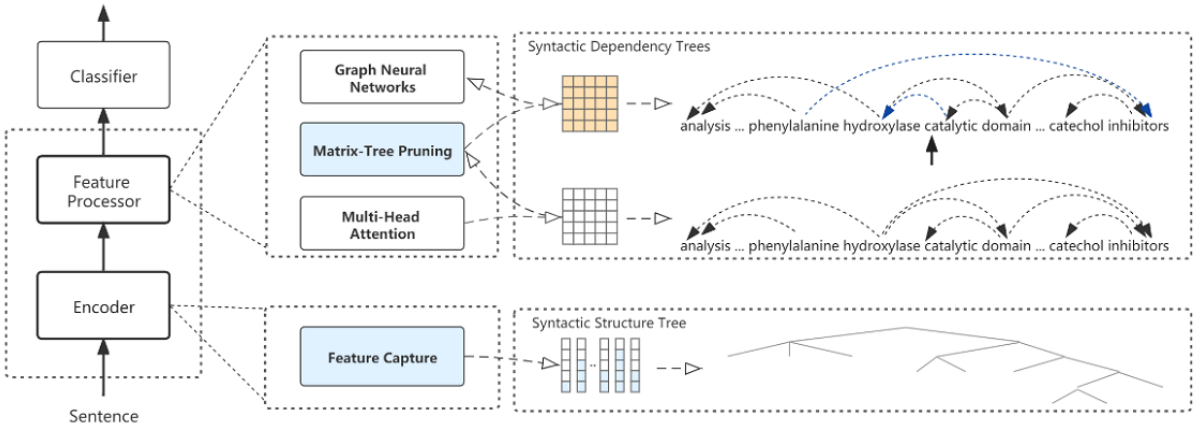
Experimental Parameter Setting

For the cross-sentence relation task, we referred to the same data divides that Guo et al [14] used. The hidden size of ON-LSTM is set to 300 in our stochastic gradient descent optimizer with a 300-dimensional Glove and 0.9 decay rate and reports the average test accuracy over 5 cross-validation folds. For the sentence-level task, the F1 results are shown [8], and we randomly divided 10% of the PGR training set as the development set to ensure consistent data division. We fine-tuned the hyperparameters based on the outcomes of the development sets. The results marked with an asterisk are based on a reimplementation of the original model. The aforementioned configuration ensures that our model has a consistent data partitioning and operating environment with the baseline.

The Overall Architecture

An overview of our proposed syntactic enhancement graph convolutional network (SEGCN) model (Figure 2) consists of 3 parts: an Encoder, a Feature Processor, and a classifier. The Encoder incorporates the syntactic structural features, and the Feature Processor handles the features containing structural information.

Figure 2. Diagrammatic representation of the syntactic enhancement graph convolutional network model showing an instance and its syntactic information processing flow. The syntactic structure tree can be obtained from the encoder, and a matrix-tree can transform the syntactic dependency tree in the feature processor.



Encoder

We used ON-LSTM [13] to obtain a syntactic structure in shallow syntactic information. The ON-LSTM introduces syntactic structure information while encoding by layering the neurons. In terms of the overall framework, it is similar to LSTM. Here, we mathematically illustrate how ON-LSTM incorporates syntactic structural features.

Given a sentence $s = x_1,\ldots,x_n$, where x_i represents the i-th word. We have written $\mathbf{h} = \mathbf{h}_1,\ldots,\mathbf{h}_n$ for the structural output of the sentence $\mathbf{h}$ $R^{n\times d}$, where $\mathbf{h}_i$ R^d denotes the i-th word’s hidden state with a d dimension. A cell c_t is used to record the state of $\mathbf{h}_t$; to control $\mathbf{h}_t$, which is the data flow between the inputs and outputs, a forget gate f_t, an output gate o_t and an input gate i_t are employed. Where $\mathbf{W}_x$, $\mathbf{U}_x$, and b_x(x f, I, o, c) are model parameters, and c_0 is a zero-filled vector:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad \textbf{(1)}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad \textbf{(2)}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad \textbf{(3)}$$

$$c_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad \textbf{(4)}$$

$$h_t = o_t \bullet \tanh(c_t) \quad \textbf{(5)}$$

It differs from the LSTM in that it uses a new function to replace the update function of the cell state c_t. Specific ordering of internal neurons by replacing the update function, allowing the syntactic structure to be integrated into the LSTM. The update rules are as follows.

(6)

(7)

(8)

We used softmax to predict the layer order of neurons and then calculate the cumulative sum by cs. Finally, f_t and i_t contains the layer order information of c_{t-1} and c_t, respectively, and the intersection of the two is ω_t. The cumulative sum equation is as follows.

 **(9)**

 **(10)**

Following the cumulative sum's properties, the master forget gate f_t has values that change from 0 to 1, while the master input gate i_t has values that decrease monotonically from 1 to 0. The overlap of f_t and i_t is represented by the product of the two master gates ω_t.

$$C = \omega_t \bullet (f_t \bullet c_{t-1} + i_t \bullet c_t) + (f_t - \omega_t) \bullet c_{t-1} + (i_t - \omega_t) \bullet c_t \quad \textbf{(11)}$$

Finally, the cell state C is segmented by layer order information, and the fused syntactic structure is fused in the model.

Feature Processor

Multi-Head Attention

By building an attention adjacency matrix $\boldsymbol{S}^k$, we converted the feature **h** to a fully connected weight graph. A set of key-value pairs and a query were used in the calculation. The obtained attention matrices represent the potential syntactic tree, which is computed from the function of the keyword **K** with the corresponding query **Q**. In this case, both **Q** and **K** are the same as **h**.

 **(12)**

Where $\mathbf{W}^Q$  $R^{d \times d}$ and $\mathbf{W}^K$  $R^{d \times d}$ are parameters for projections, d denotes the vector dimension. $\mathbf{S}^k$ consists of  . $\mathbf{h}_i$ and $\mathbf{h}_j$ represent the normalized weight scores of the i-th and the j-th token, respectively.

Matrix-Tree Pruning

We pruned the matrix-tree $\mathbf{S}^k$ to balance the syntactic features, output as matrix-tree A. It is achieved by multiplying a Gaussian kernel with an attention matrix. In the field of image processing, Gaussian kernel functions are commonly used to equalize images. In the model, we chose a 2-dimensional Gaussian kernel to balance the syntactic features. The following is the Gaussian kernel function.

 **(13)**

where a is the amplitude, x_o and y_o are the coordinates of the center point, and σ_x and σ_y are the variance. With the aforementioned 2-dimensional Gaussian kernel function, we could obtain the Gaussian kernel.

GCN

GCN is a neural network that can use information about the graph's structure. On the input of the GCN, we replaced the graph structure of the input with the syntactic tree matrix A generated above, and the feature vector is the output vector **h** of the Encoder. The layer-wise propagation rules of GCN are as follows:

 **(14)**

The adjacency matrix of an undirected graph **g** with extra self-connections is denoted by $\tilde{\mathbf{A}}$, $\tilde{\mathbf{A}} = \mathbf{A} + I_N$. I_N is the identity matrix, $D_{ii} = \Sigma_i \tilde{\mathbf{A}}_{ij}$. $\mathbf{W}^{(l)}$ is a trainable weight matrix. The activation function is denoted by $\sigma(\bullet)$. $\mathbf{H}^{(l)}$  $R^{N \times D}$ is the activation matrix in the l-th layer, $\mathbf{H}^{(0)}$ denotes the **h**.

Classifier

To obtain final categorization representations, we combined sentence and entity representations and fed them into a feedforward neural network.

$$H_{\text{final}} = FFNN([H_{sent} ; H_s ; H_o]) \quad \textbf{(15)}$$

$\mathbf{H}_{\text{sent}}$, $\mathbf{H}_s$, and $\mathbf{H}_o$ denote sentence, subject, and object representations, respectively. Finally, the logistic regression classifier performs predicted categorization of the outcome using $\mathbf{H}_{\text{final}}$ as a token.

Results

Results of the Cross-sentence Task

For the cross-sentence task, we used 3 types of models as baselines: (1) feature-based classifier [15] based on all entity pairs' shortest dependency pathways; (2) graph-structured LSTM methods, including bidirectional directed acyclic graph (DAG) LSTM (Bidir DAG LSTM) [5], Graph State LSTM (GS LSTM), and Graph LSTM [4]—these approaches extend LSTM to encode graphs generated from dependency edges created from input phrases; and (3) pruned GCNs [6] including attention-guided GCN (AGGCN) [14] and Lévy Flights GCN (LFGCN) [11]. These methods use GCNs to prune graphs with dependency edges. Additionally, we added the Bidirectional Encoder Representations from Transformers (BERT) pretraining model to complement the model with experiments. The results marked with an asterisk are based on a reimplementation of the original model.

In the multi-class relation extraction task (last 2 columns in Table 1), our SEGCN model outperforms all baselines with accuracies of 81.7 and 80.2 on all instances (Cross). In the ternary and binary relations, our SEGCN model outperforms the best performing graph-structured LSTM model (GS LSTM) by 10.0 and 8.5 points, respectively, our model outperforms the best performing model with LFGCN by 1.8 and 2.6 points when compared to the GCN models.

Table 1. Results of the cross-sentence task.

Model	Binary-class, accuracy				Multi-class, accuracy	
	Ternary		Binary		Ternary	Binary
	Single	Cross	Single	Cross	Cross	Cross
Feature-Based	74.7	77.7	73.9	75.2	__[a]	—
Graph LSTM[b]	77.9	80.7	75.6	76.7	—	—
DAG[c] LSTM	77.9	80.7	74.3	76.5	—	—
GS LSTM[d]	80.3	83.2	83.5	83.6	71.7	71.7
GCN[e] + Pruned	85.8	85.8	83.8	83.7	78.1	73.6
AGGCN[f]	87.1	87.0	85.2	85.6	80.2	77.4
LFGCN[g]	87.3	86.5	86.7	85.7	79.9	77.6
AGGCN + BERT[h]	87.2	87.1	86.1	84.9	80.5	78.1
LFGCN + BERT	87.3	86.5	86.5	86.7	80.3	78.0
SEGCN[i]	88.5	88.2	87.2	87.5	81.7	80.2
SEGCN + BERT	88.7	88.4	86.8	87.7	81.9	80.4

[a]Not determined.

[b]LSTM: long short-term memory.

[c]DAG: directed acyclic graph.

[d]GS LSTM: graph-structured long short-term memory.

[e]GCN: graph convolutional network.

[f]AGGCN: attention-guided graph convolutional network.

[g]LFGCN: Lévy Flights graph convolutional network.

[h]BERT: Bidirectional Encoder Representations from Transformers.

[i]SEGCN: syntactic edge-enhanced graph convolutional network.

In the binary-class relation extraction task, our SEGCN model also outperforms all baselines (first four columns in Table 1). The task was expanded to cross-sentence– (Cross) and sentence-level (Single) subtasks. In cross-sentence–level ternary and binary classification, our model received 88.2 and 87.5 points, respectively. Our model received 88.5 and 87.2 for sentence-level ternary and binary classifications, respectively.

These experiments show that our model achieves better results than previous models that discard shallow syntactic information, such as the previous GS LSTM and GCN models. We attribute the results of our models to the introduction of shallow syntactic information and the equalization process. Finally, for comparison with the latest methods, we attempted to introduce BERT pretraining. We found that the results of the task improved slightly after BERT pretraining. We believe that BERT also captured some shallow syntactic information during pretraining.

Results of the Sentence-Level Task

The results of the sentence-level task using the CPR [11] and PGR [16] data sets are shown in Table . Our model has been compared to 2 types of models: (1) sequence-based models, including the randomly initialized Dilated and Depthwise separable convolutional neural network (Random-DDCNN) [9], which uses a parser that is a relational prediction model through random initialization and fine-tuning; attention-based multilayer gated recurrent unit [17], which overlays attentional mechanisms on top of the recursive gated units; Bran [18], which uses a bi-affine self-attention model to capture the sentence's interactions; and Bidirectional Encoder Representations from Transformers for Biomedical Text Mining [19], which is a pretrained language representation model for medical literature; and (2) dependency-based models, which are based on a single dependency tree, including the biological ontology–based long short-term memory network [20] and GCN. There are also dependency forest–based models, including the Edgewise–graph recurrent network (GRN) [8], which prunes scores greater than a threshold; kBest-GRN [8], which involves merging of k-best trees for prediction; ForestFT-DDCNN [9], which constructs a learnable dependency analyzer; and AGGCN and LFGCN [11], which relate multiheaded attention to dependency features.

Table 2. Results of the sentence-level task.

Type and model	Multi-class (BioCreative ViCPR data set), F1 score	Binary-class (Phenotype-Gene Relationship data set), F1 score
Sequence-based model		
Random-DDCNN[a]	45.4	—[b]
Att-GRU[c]	49.5	—
Bran	50.8	—
BioBERT[d]	—	67.2
Dependency-based model		
BO-LSTM[e]	—	52.3
GCN[f]	52.2	81.3
Edgewise-GRN[g]	53.4	83.6
kBest-GRN	52.4	85.7
ForestFT-DDCNN	55.7	89.3
AGGCN[h]	56.7	88.5
LFGCN[i]	64.0	89.6
LFGCN+BERT	64.2	89.8
Our models		
SEGCN[j]	65.4	91.3
SEGCN+BERT	65.6	91.5

[a]DDCNN: Dilated and Depthwise separable convolutional neural network.

[b]Not determined.

[c]Att-GRU: attention-based multilayer gated recurrent unit.

[d]BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining.

[e]BO-LSTM: biological ontology–based long short-term memory.

[f]GCN: graph convolutional network.

[g]GRN: graph recurrent network.

[h]AGGCN: attention-guided graph convolutional network.

[i]LFGCN: Lévy Flights graph convolutional network.

[j]SEGCN: syntactic enhancement graph convolutional network.

As shown in the results of the sentence-level task in Table 2, our model achieved the best performance on both the multiclass data set CPR and the dichotomous data set PGR, with F1 scores of 65.4 and 91.3. Specifically, our model outperformed the previous state-of-the-art dependency-based model (LFGCN) by 1.2 and 1.5 points on the CPR and PGR data sets, respectively. We found that the model's improvement was smaller than that on the cross-sentence level task. We argue that shallow syntactic information has a smaller impact on short sentence lengths in sentence-level tasks, and it is better suited to long sentence lengths in cross-sentence tasks.

Discussion

Ablation Study

We validated the different modules of our model on the PGR data set, including BERT pretraining, the matrix-tree pruning layer, and the feature capture layer. Table 3 shows these results. We can see that model effectiveness decreases after removing any of the modules. All three modules can aid in the model's learning of a more accurate feature representation. The feature capture layer and the matrix-tree pruning layer improved by 2.4 and 2.5 points, respectively, indicating that the shallow syntactic information and equalization process resulted in a model boost. In contrast, the popular BERT pretraining approach was not suitable for the model.

Table 3. An ablation study using the Phenotype-Gene Relationship data set.

Model	F1 score
SEGCN[a] (All)	91.5
SEGCN (- BERT Pretraining)	91.3
SEGCN (- Matrix-tree pruning)	90.0
SEGCN (- Feature capture)	89.1
Baseline (- All)	88.5

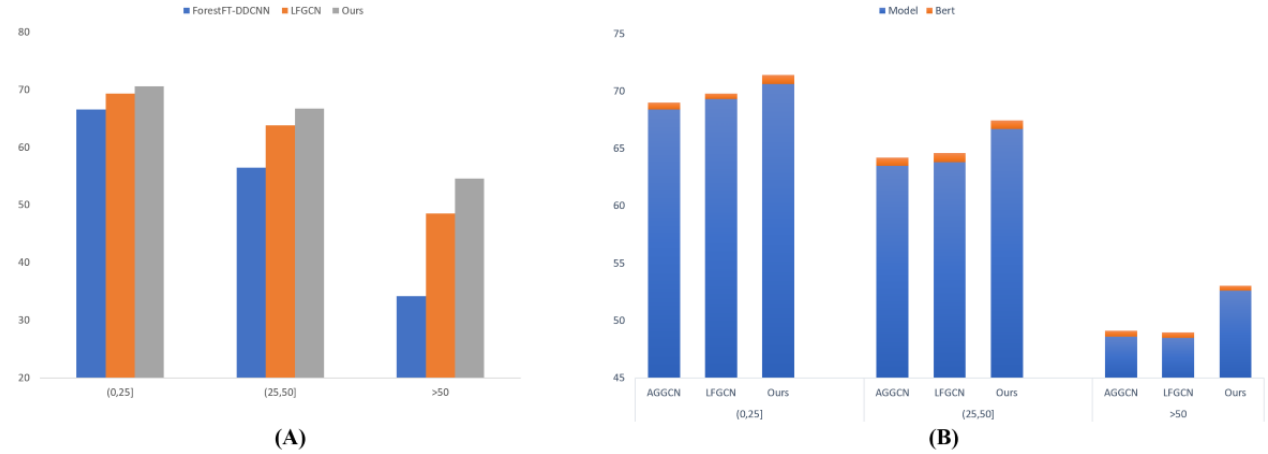
[a]SEGCN: syntactic enhancement graph convolutional network.

The ablation experiments show that shallow syntactic information and equalization processing methods can improve model performance significantly. We believe that these two methods function by processing the interaction information in the sentences. The shallow syntactic information complements the nonlocal interaction of the sentence, and the equalization process balances the local and nonlocal interactions of the sentence.

Performance Against Sentence Length

We examined the effect of introducing shallow syntactic information on different sentence lengths through comparative experiments. Figure 3A shows the F1 scores of the 3 models at different sentence lengths. There are 3 categories based on sentence length ((0,25), [25,50),>50). In general, our SEGCN outperformed ForestFT-DDCNN and LFGCN in all 3 length categories. Furthermore, the performance gap widened as the instance length increased. These results suggest that adding shallow syntactic information, particularly in long sentences, improves our model significantly. We attribute this to the fact that our model complements the nonlocal interactions of the sentences with the introduction of shallow syntactic information. Because they rely more on nonlocal interactions, longer sentences received higher F1 scores.

Figure 3. Performance against sentence length and Bidirectional Encoder Representations from Transformers (BERT) pretraining. (A) F1 scores at different sentence lengths. Results of the ForestFT– Dilated and Depthwise separable convolutional neural network are based on Jin et al [10]. (B) F1 scores against sentence length after BERT pretraining. AGGCN: attention-guided graph convolutional network; LFGCN: Lévy Flights graph convolutional network.



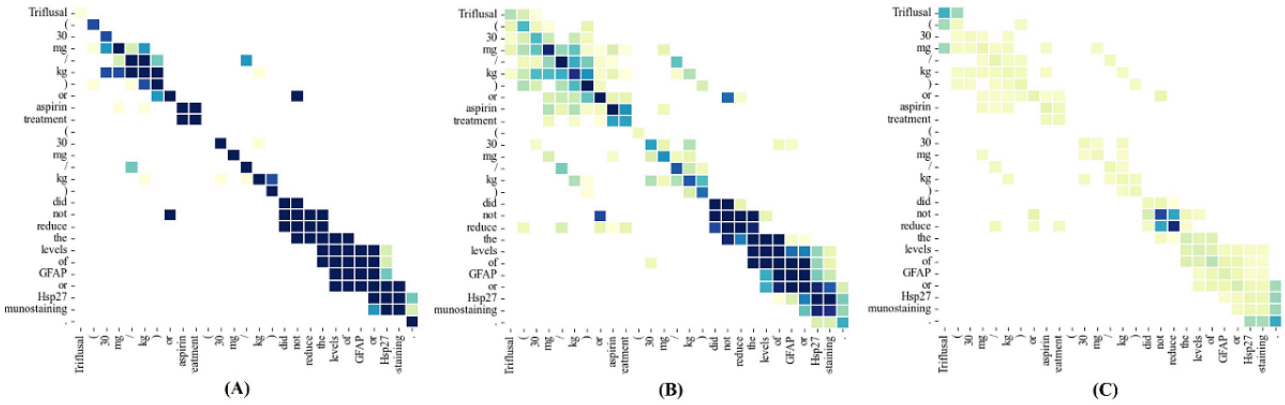
Performance Against BERT Pretraining

To show the superiority of syntactic enhancement of our models, we compared the models with the addition of pretraining. After BERT pretraining, the F1 scores of the 3 models are shown in Figure 3B for different sentence lengths. There are 3 categories based on sentence length ((0,25], [25,50),>50). Overall, BERT pretraining showed small improvements for models of different sentence lengths. It supports our hypothesis that the neural network models acquire insufficient syntactic features. Furthermore, we found that our SEGCN without BERT still functioned better than the other models with BERT. These results indicate that our model outperforms BERT in using syntactical features.

Case Study

To demonstrate the impact of our approach on sentence interaction, we compared the features obtained from different model layers. Figure 4 shows the attention weights of the example sentences at the different layers of the model. We decided to use a heat map to represent the attention weights. The color of each point represents the weight of the interactive information. The darker the color, the greater the weighting. For more intuition, we have omitted the points with smaller weights. In addition, the output of the multi-headed attention layer before and after incorporation into the shallow syntactic information is represented by matrices A and B, respectively. Matrix C represents the output of the equalization processing matrix B.

Figure 4. The heat maps of an example sentence in the syntactic enhancement graph convolutional network model.



As shown in Figure 4, the weight distribution in matrix **A** is more concentrated in the diagonal distribution. In contrast, matrix B and matrix C have significantly more nondiagonal weight distributions than matrix A. This supports our view that the model incorporating shallow syntactic information gradually focuses on nonlocal interactions in the sentence. Furthermore, by comparing matrices B and C, we see that equalized matrix C pays more even-handed attention to the model's weights (the more similar the color, the closer the weights). We believe that the model's performance is improved by balancing the attention to local and nonlocal interactions. These results further demonstrate how our model makes use of syntactic information for syntactic enhancement.

Conclusions

This study is the first to propose incorporating shallow syntactic information for syntactic enhancement in medical relation extraction. In addition, we devised a new pruning method to equalize the syntactic interactions in the model. The results for the 3 medical data sets show that our method can improve and equalize syntactic interactions, significantly outperforming previous models. The ablation experiments demonstrate the effectiveness of our two proposed methods. In future, we intend to continue our research on the connection between shallow syntactic information and sentence interactions.

Acknowledgments

The publication of this paper is funded by grants from the Natural Science Foundation of China (62006034 and 62072070), Natural Science Foundation of Liaoning Province (2021-BS-067), and the Fundamental Research Funds for the Central Universities [DUT21RC (3)015].

Authors' Contributions

WT led the method application, experiment conduction, and the result analysis. DZ participated in the data extraction and preprocessing. YZ participated in the manuscript revision. HM provided theoretical guidance and the revision of this paper.

Conflicts of Interest

None declared.

References

1. Heeman PA, Allen JF. Incorporating POS Tagging Into Language Modeling. 1997 Presented at: Fifth European Conference on Speech Communication and Technology, EUROSPEECH; September 22-25, 1997; Rhodes URL: https://www.cs.rochester.edu/research/cisd/pubs/1997/paper1.pdf
2. Wright JH, Jones GJF, Lloyd-Thomas H. A robust language model incorporating a substring parser and extended n-grams. 1994 Presented at: ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing; April 19-22, 1994; Adelaide, SA. [doi: 10.1109/icassp.1994.389281]
3. Merity S, Keskar NS, Socher R. Regularizing and optimizing LSTM language models. 2018 Presented at: 6th International Conference on Learning Representations, ICLR 2018; April 30 - May 3, 2018; Vancouver, BC.
4. Peng N, Poon H, Quirk C, Toutanova K, Yih W. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. TACL 2017 Dec;5:101-115. [doi: 10.1162/tacl_a_00049]
5. Linfeng S, Yue Z, Zhiguo W. N-ary Relation Extraction using Graph-State LSTM. 2018 Presented at: 2018 Conference on Empirical Methods in Natural Language Processing; October 31, 2018; Brussels. [doi: 10.18653/v1/d18-1246]

6. Zhang Y, Qi P, Manning CD. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. 2018 Presented at: 2018 Conference on Empirical Methods in Natural Language Processing; October 31, 2018; Brussels. [doi: 10.18653/v1/d18-1244]
7. Zhang H, Lu G, Zhan M, Zhang B. Semi-Supervised Classification of Graph Convolutional Networks with Laplacian Rank Constraints. Neural Process Lett 2021 Jan 01. [doi: 10.1007/s11063-020-10404-7]
8. Song L, Zhang Y, Gildea D, Yu M, Wang Z, Su J. Leveraging Dependency Forest for Neural Medical Relation Extraction. 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 2019; Hong Kong. [doi: 10.18653/v1/d19-1020]
9. Jin L, Song L, Zhang Y, Xu K, Ma W, Yu D. Relation Extraction Exploiting Full Dependency Forests. 2020 Apr 03 Presented at: AAAI Conference on Artificial Intelligence; February 7–12, 2020; New York, NY. [doi: 10.1609/aaai.v34i05.6313]
10. Dozat T, Manning CM. Deep biaffine attention for neural dependency parsing. 2017 Presented at: 5th International Conference on Learning Representations, ICLR 2017; April 24-26, 2017; Toulon.
11. Zhijiang G, Nan G, Lu W, Cohen SB. Learning Latent Forests for Medical Relation Extraction. 2020 Presented at: Twenty-Ninth International Joint Conference on Artificial Intelligence; 2020; Yokohama. [doi: 10.24963/ijcai.2020/505]
12. Hale J, Dyer C, Kuncoro A, Brennan J. Finding syntax in human encephalography with beam search. 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; July 2018; Melbourne, VIC. [doi: 10.18653/v1/p18-1254]
13. Shen Y, Tan S, Sordoni A, Courville A. Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks. arXiv. Preprint posted online May 8, 2019 2019.
14. Guo Z, Zhang Y, Lu W. Attention Guided Graph Convolutional Networks for Relation Extraction. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence. [doi: 10.18653/v1/p19-1024]
15. Quirk C, Poon H. Distant Supervision for Relation Extraction beyond the Sentence Boundary. 2017 Presented at: 15th Conference of the European Chapter of the Association for Computational Linguistics; April 2017; Valencia. [doi: 10.18653/v1/e17-1110]
16. Sousa D, Lamurias A, Couto FM. A Silver Standard Corpus of Human Phenotype-Gene Relations. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2019; Minneapolis, MN. [doi: 10.18653/v1/n19-1152]
17. Liu S, Shen F, Komandur Elayavilli R, Wang Y, Rastegar-Mojarad M, Chaudhary V, et al. Extracting chemical-protein relations using attention-based neural networks. Database (Oxford) 2018 Jan 01;2018:102 [FREE Full text] [doi: 10.1093/database/bay102] [Medline: 30295724]
18. Verga P, Strubell E, McCallum A. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2018; New Orleans, LA. [doi: 10.18653/v1/n18-1080]
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
20. Lamurias A, Sousa D, Clarke LA, Couto FM. BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. BMC Bioinformatics 2019 Jan 07;20(1):10 [FREE Full text] [doi: 10.1186/s12859-018-2584-5] [Medline: 30616557]

Abbreviations

AGGCN: attention-guided graph convolutional network
BERT: Bidirectional Encoder Representations from Transformers
DAG: directed acyclic graph
DDCNN: Dilated and Depthwise separable convolutional neural network
GCN: graph convolutional network
GRN: graph recurrent network
LFGCN: Lévy Flights graph convolutional network
LSTM: long short-term memory
ON-LSTM: ordered neuron–long short-term memory
PGR: Phenotype-Gene Relationship
Random-DDCNN: randomly initialized Dilated and Depthwise separable convolutional neural network
SEGCN: syntactic enhancement graph convolutional network

Edited by T Hao; submitted 10.03.22; peer-reviewed by J Gao, Y Du; comments to author 28.05.22; revised version received 01.06.22; accepted 27.06.22; published 02.08.22.

Please cite as:
Tang W, Wang J, Lin H, Zhao D, Xu B, Zhang Y, Yang Z
A Syntactic Information–Based Classification Model for Medical Literature: Algorithm Development and Validation Study
JMIR Med Inform 2022;10(8):e37817
URL: https://medinform.jmir.org/2022/8/e37817
doi:10.2196/37817
PMID:35917162

Original Paper

Emotion-Based Reinforcement Attention Network for Depression Detection on Social Media: Algorithm Development and Validation

Bin Cui[1], MSc; Jian Wang[1], PhD; Hongfei Lin[1], PhD; Yijia Zhang[2], PhD; Liang Yang[1], PhD; Bo Xu[1], PhD

[1]College of Computer Science and Technology, Dalian University of Technology, Dalian, China

[2]College of Information Science and Technology, Dalian Maritime University, Dalian, China

Corresponding Author:
Jian Wang, PhD
College of Computer Science and Technology
Dalian University of Technology
Number 2, Linggong Road
Ganjingzi District
Dalian, Liaoning 116024
China
Phone: 86 13604119266
Email: wangjian@dlut.edu.cn

Abstract

Background: Depression detection has recently received attention in the field of natural language processing. The task aims to detect users with depression based on their historical posts on social media. However, existing studies in this area use the entire historical posts of the users and select depression indicator posts. Moreover, these methods fail to effectively extract deep emotional semantic features or simply concatenate emotional representation. To solve this problem, we propose a model to extract deep emotional semantic features and select depression indicator posts based on the emotional states.

Objective: This study aims to develop an emotion-based reinforcement attention network for depression detection of users on social media.

Methods: The proposed model is composed of 2 components: the emotion extraction network, which is used to capture deep emotional semantic information, and the reinforcement learning (RL) attention network, which is used to select depression indicator posts based on the emotional states. Finally, we concatenated the output of these 2 parts and send them to the classification layer for depression detection.

Results: Experimental results of our model on the multimodal depression data set outperform the state-of-the-art baselines. Specifically, the proposed model achieved accuracy, precision, recall, and F1-score of 90.6%, 91.2%, 89.7%, and 90.4%, respectively.

Conclusions: The proposed model utilizes historical posts of users to effectively identify users' depression tendencies. The experimental results show that the emotion extraction network and the RL selection layer based on emotional states can effectively improve the accuracy of detection. In addition, sentence-level attention layer can capture core posts.

(JMIR Med Inform 2022;10(8):e37818) doi:10.2196/37818

KEYWORDS

depression detection; emotional semantic features; social media; sentence-level attention; emotion-based reinforcement

Introduction

As an important part of medical informatics research, depression is one of the most dangerous diseases impacting human mental health. It is different from usual mood swings and transient emotional reactions. Long-term depression may cause severe problems for the patient, such as suicide. The World Health Organization (WHO) ranks depression as the most significant cause of disability [1]. Statistics show that over 300 million people suffer from depression all over the world, and the number of patients continues to grow [2]. Depression detection for potential users can help detect the disease at an early stage and help patients get timely treatment.

The latest global digital report [3] shows that there are 4.62 billion social media users worldwide, which is equivalent to 58.4% of the world's population. Internet users worldwide spend

nearly 7 hours a day on the web and 2 hours and 30 minutes on social media. Over the past year, social media users have increased by an average of more than 1 million per day. All these show that social media plays a central role in our daily lives. Meanwhile, an increasing number of people tend to express their emotions and feelings on Weibo, Twitter, etc. People with depression are willing to post depression-related information on social media, such as negative emotions or depression treatment information [4,5]. Therefore, we can obtain a great deal of valuable information about depression from their tweets. The objective of this paper is to predict a label {depression, nondepression} for each user indicating their depressive tendencies by mining their historical posts.

In recent years, psychology-related social media mining has become a research hotspot in natural language processing. The task of detecting users with depression through historical posts on social media has received extensive attention from researchers. Many computer researchers and psychologists have proposed effective methods to detect depression by extracting emotion, interaction, and other features from texts. Nguyen et al [6] extracted emotions, psycholinguistic processes, and content themes in posts to detect users with depression. Shen et al [7] constructed well-labeled depression data sets on Twitter and extracted 6 feature groups associated with depression. Tong et al [8] extracted 3 discriminative features from users' posts, and then proposed a new cost-sensitive boosting pruning trees model to detect users with depression. Park et al [9] concluded that users with depression prefer to express their status on social media than in real life, so extracting emotional information was essential for depression-detection tasks.

With the maturity of deep learning, the research models have gradually moved from traditional feature engineering to deep learning methods. Yates et al [10] utilized a convolutional neural network (CNN)–based model with multiple inputs for detecting users with depression. Alhanai et al [11] used long short-term memory network (LSTM) to concatenate text and audio representation to detect users with depression. Ren et al [12] extracted emotional information by combining positive words and negative words. Orabi et al [13] investigated the performance differences between recurrent neural network (RNN) models and CNN models in depression detection. Zogan et al [14] fused semantic and user behavior information for detecting depression, and proposed the multimodal depression detection with hierarchical attention network (MDHAN).

All these aforementioned deep learning methods use the entire historical posts of the users. However, it is common for users to share various posts online, and posts related to depression are usually rare. The large number of irrelevant posts contained in historical posts can degrade the performance of the model. Figure 1 illustrates this phenomenon, where posts related to depression are highlighted in red, and the irrelevant posts are highlighted in blue.

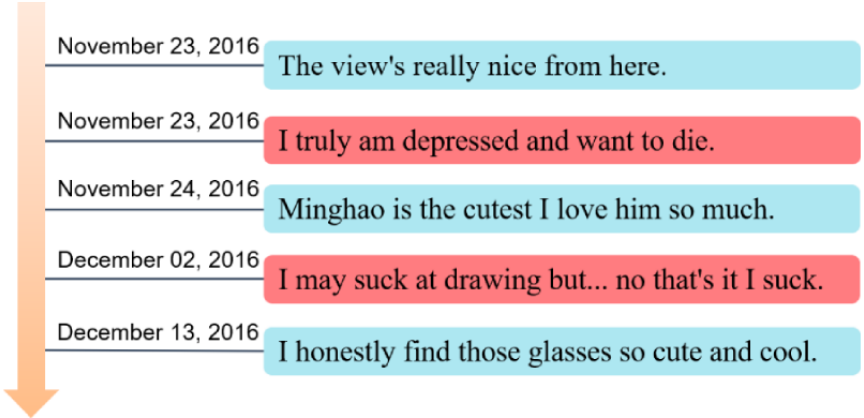
From Figure 1, we can see that only a small percentage of tweets are related to depression. Gui et al [15] selected depression indicator posts by reinforcement learning (RL). The advantage of selecting indicator posts is that it excludes the influence of irrelevant posts. If we take all the user's posts as input, a large amount of noise will be introduced.

From this example, we can also see that there are many emotional words in the user's posts such as "depressed", "suck", "die", "nice". However, current methods are lacking in deep mining of emotional information and do not well integrate emotional information into the model. Motivated by these, we propose an emotion-based reinforcement attention network (ERAN) for depression detection in this paper. The proposed model effectively improves the accuracy of depression detection by extracting deep emotional features, selecting depression indicator posts based on the current emotional states, and capturing core information through the sentence-level attention.

The main contributions of this paper can be summarized in the following 3 points:

- First, we extract emotional features by the pretrained TextCNN and fuse the emotional vectors with the output of the attention layer to classify users.
- Second, we improve a reinforcement attention network, which is mainly composed of an RL selection layer and a sentence-level attention layer. The RL selection layer can select depression indicator posts based on the emotional states, and the sentence-level attention captures core information by assigning different weights to posts.
- Finally, experimental results show that the proposed model outperforms the state-of-the-art baselines on the multimodal depression data set (MDD).

Figure 1. Sample posts of a depressed user. The posts with red highlights are considered depression indicator posts.



Methods

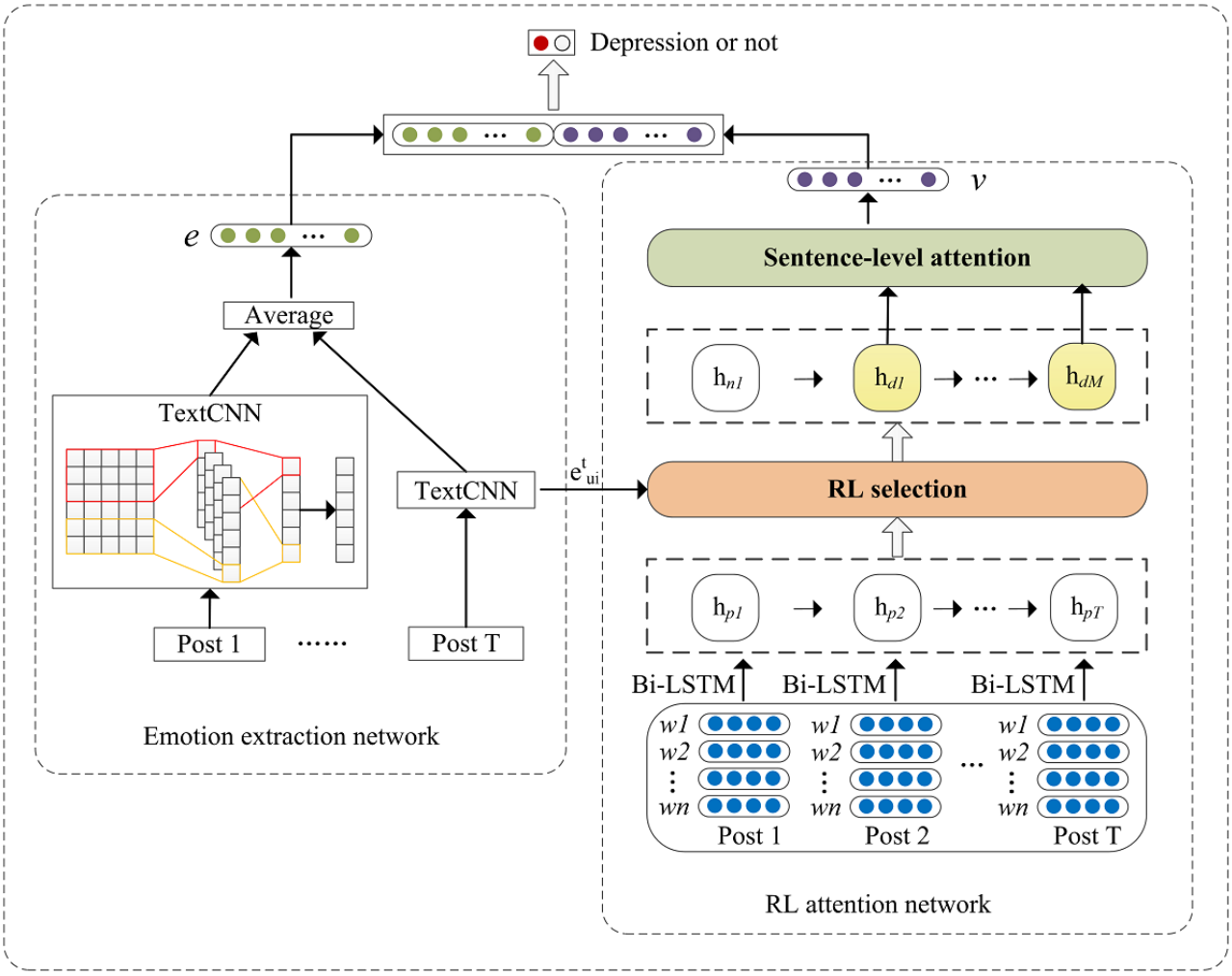
Task Definition

Let $H_i = \{p^1_i, p^2_i, ..., p^T_i\}$ be the set of T historical posts of user u_i. The goal of the depression detection is to predict a label to the user u_i based on historical posts to indicate whether the user is depressed or not.

Model Overview

In the following, we will introduce the structure of our model for depression detection. The proposed model consists of 2 networks, including an emotion extraction network and an RL attention network. The emotion extraction network is used to capture deep emotional sentiment representation from a user's historical posts. The RL attention network selects depression indicator posts based on the emotional states and assigns weights for the selected posts by the sentence-level attention. Finally, we concatenate the representations captured by the 2 networks and send them to the classification layer to detect whether the user is depressed or not. Figure 2 shows the architecture of the proposed model.

Figure 2. Architecture of the Emotion-Based Reinforcement Attention Network (ERAN). LSTM: long short-term memory network; RL: reinforcement learning.



Emotion Extraction Network

Many studies have shown that emotional information is essential for depression detection on social media. However, current methods fail to extract deep emotional semantic information effectively or do not incorporate the emotional representation well into the model. For instance, some methods just simply concatenate sentiment representation with other information. Motivated by this, we used a pretrained TextCNN [16] to extract deep sentiment features and feed them to the RL attention network of the proposed model to accomplish deep interactions.

For user u_i, we input all posts p^t_i into a pretrained TextCNN. The TextCNN has been pretrained on an emotion classification task labeled as positive, negative, and neutral. After training, the TextCNN is used to extract the emotional information of each post. We regard the last hidden layer vector of the TextCNN as emotion vector . The final emotional semantic representation for all T-posts of user u_i is defined as , which is the expectation of :

where T is the number of posts by user u_i and t is tth post of the user u_i.

Let  denote the representation of a user's post, with n as the length of the padded post.  represents the concatenation operator. We utilize word2vec [17] to encode each word w_i as a d-dimensional word embedding  .

Then, we input the text sequence $X_{1:n}$ into a single-layer CNN. The convolutional layer of the CNN has 3 filters  . For each  , there is Z filter F_k for extracting complementary information. And then, we apply them to a window  to generate a new feature vector. The feature vector $c_{k,j}$ is calculated by:

where $\alpha(\cdot)$denotes a nonlinear activation function;  is a window with h_k words, and  is a bias. For each window in the post $\{X_{1:h}, X_{2:h+1}, \ldots, X_{n-h+1:n}\}$, the above actions are taken to get a feature map  , where  , and h_k is the height of the convolution kernel.

After convolution operation, each filter F_k creates Z feature maps  . Following this, to extract the maximum features, we connect a max-pooling operation [18] to all feature maps. The output is calculated as  . The output of max-pooling, which covers all feature maps  , is the concatenation of each $\boldsymbol{o}_k$. Finally,  is entered into a fully connected layer. The output of the classification layer is calculated as:

where  , and  ; $\alpha(\cdot)$ is a nonlinear activation function. The fully connected layer is followed by a sigmoid-classification layer with 3 classes, and $\sigma(\cdot)$ represents sigmoid operation.

RL Attention Network

Overview

Users' historical posts usually contain various content, and only a small fraction may be related to depression. Those irrelevant posts pose a challenge to identify users' depressive tendencies effectively, so we need to develop a model to select only depression-related posts. The historical posts of the user u_i are denoted as $H_i = \{p^1_i, p^2_i, \ldots, p^T_i\}$, and the depression indicator posts are denoted as  .

The structure of this network includes (1) a bidirectional LSTM (BiLSTM) that generates contextual representation, (2) an RL selection layer that chooses depression-related posts based on the current emotional states from H_i, and (3) a sentence-level attention layer that allows the model to pay more attention to higher-weight posts.

BiLSTM Layer

Graves et al [19] proposed the BiLSTM, which has been widely used in natural language processing to capture long-distance contextual dependency. Superior to LSTM [20], BiLSTM can capture bidirectional semantic dependencies. Inspired by this, we utilized BiLSTM to encode contextual information. The algorithm processes of LSTM are as follows:

$$f_k = \sigma(W^f \cdot [h_{k-1}, x_k] + b^f) \quad (4)$$

$$i_k = \sigma(W^i \cdot [h_{k-1}, x_k] + b^i) \quad (5)$$

$$o_k = \sigma(W^o \cdot [h_{k-1}, x_k] + b^o) \quad (6)$$

$$c_k = \tanh(W^c \cdot [h_{k-1}, x_k] + b^c) \quad (7)$$

where W^f, W^i, W^o, and W^c are parameters that can be trained.  represents the element-wise multiplication operation, x_k denotes the pretrained word2vec embedding, and $\sigma(\cdot)$ represents sigmoid function.

Given an input sequence $X = [x_1, x_2, ..., x_n]$, the forward hidden state is  , and the backward hidden state is  . The representation of the sentence is:

For user u_i, the representation of posts is  , where T is the number of posts.

RL Selection Layer

Because we only have user-level labels, it becomes a key challenge to select posts related to depression. Gui et al [15] utilized RL to select depression indicator posts. However, their method still has a high recognition accuracy in the unselected posts, which indicates that this model misses many important posts. Inspired by this, we introduced emotional states to improve the selection strategy based on RL.

RL is a way of learning by "trial and error" in the environment. It has 3 important factors: agent, environment, and reward, where the agent is the selector. At each step t, the agent executes the action a^t based on the state s^t to select the current post or not. After executing all posts, the classifier gives the agent a total reward to evaluate the performance of this policy. Policy gradient [21] is an optimization method of parameterizing the policy, which optimizes the parameter θ to maximize the total reward. Next, we will explain these parts.

In this layer, after encoding, the post p^t is denoted by the vector  . At each step t, the current post is  , the selected posts set is  , and the unselected posts set is  . If action a^t=1, the post

 is appended to  ; otherwise  is appended to $\boldsymbol{H}^{non}$, where  . The state s^t with emotional vector is represented as follows:

where  represents the concatenation operation, and avg(·) represents the average operation.  denotes the emotion vector of the tth post of u_i. The current state s^tincorporates the emotion vector, which enables the agent to take better actions. The action obeys the following policy to take actions:

$$\pi(a^t|s^t; \theta) = p_\theta(a^t|s^t, \theta) \textbf{ (12)}$$

where θ represents the parameter of the policy function and is optimized to maximize the total reward, $(a^t|s^t;\theta)$ represents the policy function that the agent follows to take action, and $p_\theta(a^t|s^t, \theta)$ is a probability distribution over the action, and we serialize the discrete policy via the *MLP* layer.

For each episode $\tau = \{s^1, a^1, s^2, a^2, ..., s^T, a^T, \text{END}\}$ of user u_i, the classifier will return a reward after all the selections are made. The objective is to maximize the reward of the episode. The reward is defined as the predicted probability after executing this episode:

$$R(\tau) = p(y_i|H^{\text{dep}}; \theta') \textbf{ (13)}$$

where θ′ represents the parameters of the classification layer and is optimized by the depression classifier.

After N sampling for user u_i, we get N episodes $\tau = \{\tau_1, ..., \tau_N\}$. To optimize the parameter θ, we calculate the expectation of $R(\tau)$. The calculation processes are as follows:

Here, because the transfer between states is Markovian, we will use the chain rule to calculate $p(\tau|\theta)$, as shown in Equation (15).

To maximize  , we calculate its gradient against θ. The equation is shown as follows:

Here, to simplify the objective function, we assume that the probability of each occurring is $1/N$. In the equation,  is a baseline value. If $R(\tau_n) - b$ is positive, the optimization will proceed toward increasing the probability $p(a^t|s^t, \theta)$. If $R(\tau_n) - b$ is negative, the optimization will proceed toward reducing the probability. Thus, is updated in this way:  , where α is the learning rate.

Finally, the loss function of this part is calculated by:

Here, maximizing $R(\tau)$ is minimizing $\text{loss}_1(\theta)$ actually. The parameters, as well as the loss, will be optimized by the gradient. After the selection of agent,  contains the posts related to depression. Then we feed $\boldsymbol{H}^{dep}$ into the attention layer.

The Sentence-Level Attention Layer

The semantics of a document can be described by a few sentences in the document. The model will not capture the key information if it treats each sentence fairly. To solve the document classification problem, Yang et al [22] designed the hierarchical attention network. This network contains a word-level attention used to focus on keywords and a sentence-level attention used to focus on critical sentence. Inspired by this, we utilized the sentence-level attention mechanism to enable our model to focus on relevant posts. It will create an attention weight for each post in $\boldsymbol{H}^{dep}$, and the model will focus more on tweets with higher weights.

We assume that the depression indicator posts set of u_i is  , which has M indicator posts after padding. For the vector  , the attention weight is calculated by:

where  is the final posts representation that summarizes all the posts in $\boldsymbol{H}_i^{dep}$  is a vector used to measure the weight of the posts and is randomly initialized. During the training process,  can be updated.

Final Prediction

In the classifier, we concatenate the output of attention layer  and emotion representation  to form the unified text representation  . Finally,  is projected to the output layer having 2 neurons with a soft-max activation. The categorical cross-entropy loss function and the soft-max probability are calculated as follows:

where, j represents the categories, U is the total number of users in data set,  represents the classification probability, and $\boldsymbol{y}_i^j$ is the ground truth.

Ethics Approval

The data set and methods used in this work are publicly available and do not involve any ethical or moral issues.

Results

Data Sets

Shen et al [7] proposed the MDD data sets, which contain well-labeled data sets D_1, D_2, and an unlabeled data set D_3 on Twitter. These 3 data sets collect posts from users on Twitter at specific times. Table 1 describes the statistics of these 3 data sets, including the number of users and tweets.

- Depression data set *D1*: Based on the tweets between 2009 and 2016, if users' tweets satisfy the strict pattern "(I'm/ I was/ I am/ I've been) diagnosed depression," they will be labeled as depressed.
- Nondepressed data set *D2*: In this data set, only users who have never posted tweets containing "depress" are marked as nondepressed.
- Depression-candidate data set *D3*: In this data set, users are obtained if their anchor tweets loosely contain "depress." In this way, *D3* contains more users with depression than randomly sampling.

In our experiments, we added all the users in D_1 to the data set. In addition, we randomly selected the same number of users in D_2 to balance the data set. Selection rules excluded users with less than 15 posts, or users with non-English posts. The data set used in this paper contained 2804 Twitter users and over 500,000 posts made by them. Finally, we used 2243/2804 (79.99%) users in the data set to train our model and 561/2804 (20%) users to test our model.

Table 1. Summary of the data sets.

Data set	Label	User	Tweets
D_1	Depressed	1402	292,564
D_2	Nondepressed	>300 million	>10 million
D_3	Nonlabeled	36,993	35,076,677

Evaluation Metrics

In the experimental phase, we used accuracy, precision, recall, and F_1-score to evaluate the performance of the proposed model. F_1-score is calculated as follows:

$$F_1 = (2 \cdot P \cdot R)/(P + R) \quad \textbf{(23)}$$

where $R = TP/(TP+FN)$ and $P = TP/(TP+FP)$; here, P is precision, R represents recall, TP represents true-positive prediction, FN is false-negative prediction, and FP is false-positive prediction.

Experimental Setting

During the experimental phase, the hyperparameters were randomly initialized based on our experience. The pretrained word2vec is used to initialize the word embeddings. The Adam optimizer [23] was used to optimize the hyperparameters. Other hyperparameter settings are shown in Table 2.

The training of our ERAN model is based on the operating system of Ubuntu 18.04, using PyTorch version 1.9.0 and Python version 3.7.0. The graphics processing unit is NVIDIA TITAN Xp with 12-GB memory.

Table 2. Values of hyperparameters.

Hyperparameters	Value
Word embedding dimension	300
BiLSTM[a] hidden units	200
Dropout rate	0.5
Batch size	128
Learning rate	0.001

[a]BiLSTM: bidirectional long short-term memory network.

Comparison With Existing Methods

Here, we describe the baseline methods that we compared with.

- Naïve Bayesian (NB): NB [24] is widely used in classification tasks. The classifier accepts all features to detect the user's depressive tendencies.
- Wasserstein Dictionary Learning (WDL): Rolet et al [25] proposed the WDL. It considers the Wasserstein distance as the fitting error to leverage the similarity shared by the features.
- Multiple Social Networking Learning (MSNL): Song et al [26] proposed the MSNL model to solve the volunteerism tendency prediction problem.
- Multimodal Depressive Dictionary Learning (MDL): Shen et al [7] proposed the MDL model by combining the multimodal strategy and dictionary learning strategy.
- CNN/LSTM + RL: Gui et al [15] proposed an RL model to select depression indicator posts.
- MDHAN: Zogan et al [14] proposed MDHAN. They extracted semantic information using a hierarchical attention network and user behavior by a multimodal encoder.

We compared the performance of the proposed model (ERAN) with other existing models on the MDD data set. The experimental results are shown in Table 3.

From the first 4 classic methods, MDL achieves the best performance with 78.6% in F_1-score, indicating the validity of the multimodal depressive dictionary. The results based on BiLSTM are better than those based on LSTM, indicating that the bidirectional encoder can capture more helpful information. Similarly, the performances based on BiLSTM (Att) are better than those based on BiLSTM, which can demonstrate that the sentence-level attention mechanism can capture more important depression information.

With the popularity of pretrained approaches, we experimented with 2 pretrained models, Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT pre-training Approach (RoBERTa) [27], and fine-tuned them on our data set. From Table 3, we can see that the simple pretraining models do not work very well, which may be due to the sparse distribution of depression-related words causing the pretrained models to fail to maximize their ability.

Table 3. Results compared with the baseline models.

Model	Accuracy	Precision	Recall	F_1-score
NB[a] [22]	0.636	0.724	0.623	0.588
WDL[b] [24]	0.761	0.763	0.762	0.762
MSNL[c] [25]	0.782	0.781	0.781	0.781
MDL[d] [6]	0.790	0.786	0.786	0.786
LSTM[e]	0.797	0.812	0.813	0.812
BiLSTM[f]	0.805	0.817	0.818	0.817
BiLSTM (Att[g])	0.817	0.828	0.828	0.828
BERT[h] (base) [27]	0.845	0.883	0.825	0.853
RoBERTa[i] (base) [27]	0.851	0.902	0.837	0.868
CNN[j] + RL[k] [14]	0.871	0.871	0.871	0.871
LSTM + RL [14]	0.870	0.872	0.870	0.871
MDHAN[l] [13]	0.895	0.902	0.892	0.893
ERAN[m] (ours)	0.906	0.912	0.897	0.904

[a]NB: naïve Bayesian.
[b]WDL: Wasserstein Dictionary Learning.
[c]MSNL: Multiple Social Networking Learning.
[d]MDL: Multimodal Depressive Dictionary Learning.
[e]LSTM: long short-term memory network.
[f]BiLSTM: bidirectional long short-term memory network.
[g]Att: attention.
[h]BERT: Bidirectional Encoder Representation from Transformers.
[i]RoBERTa: Robustly Optimized BERT pre-training Approach.
[j]CNN: convolutional neural network.
[k]RL: reinforcement learning.
[l]MDHAN: multimodal depression detection with hierarchical attention network.
[m]ERAN: emotion-based reinforcement attention network.

The CNN/LSTM + RL models use RL to select indicator posts, which verifies the validity of the selection strategy. The MDHAN model proves that the multimodal features are also important by fusing semantic information with user behavior information.

The proposed ERAN model achieves optimal results because we fused emotional information and selected depression indicator posts based on emotional states. In addition, the sentence-level attention can capture core posts.

Ablation Study

Ablation experiments were conducted to validate the necessity of the emotion extraction network, the RL selection layer, and the sentence-level attention. The study is performed by removing one module at a time. The results of the ablation experiments are presented in Figure 3.

Emotion-based BiLSTM attention network (EBAtt) is the model that removes the RL selection layer from the proposed model and uses all user posts. Reinforcement learning attention network

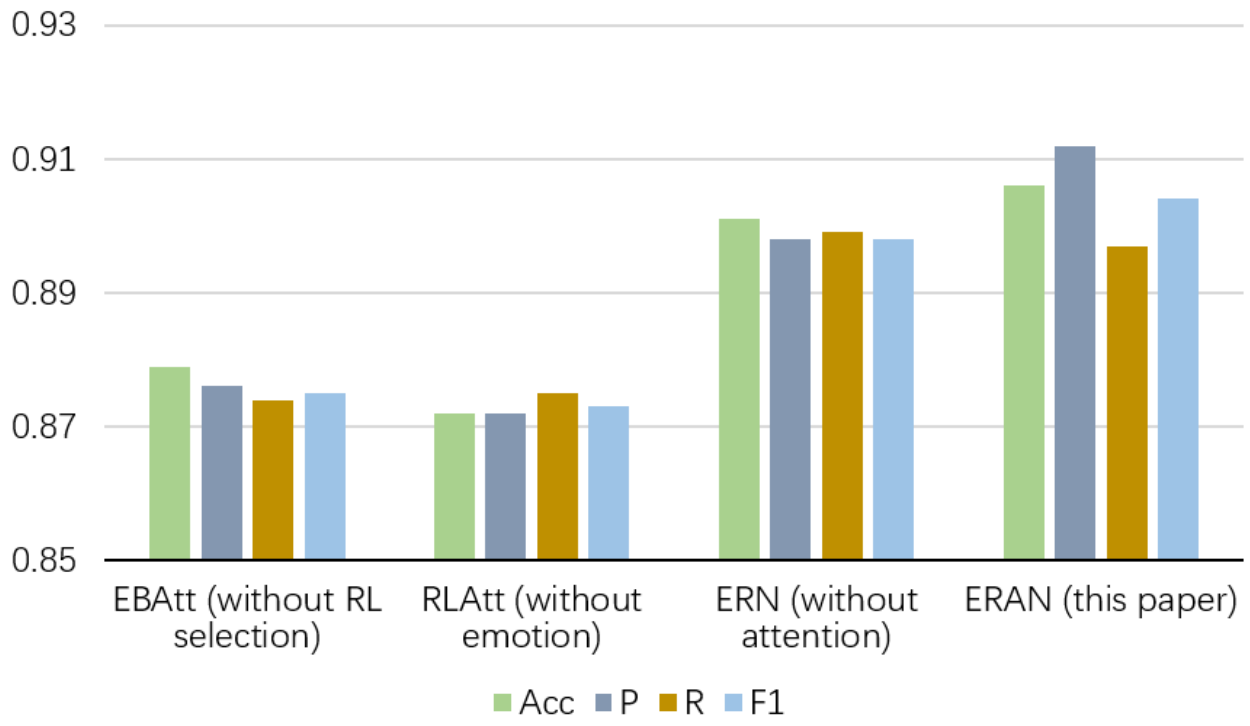
(RLAtt) is the model that removes the emotion extraction network. Emotion-based reinforcement learning network (ERN) is the model that substitutes the sentence-level attention with the averaging operation. We can see that the ERAN model proposed in this paper performs best. Although ERAN is lower than ERN in precision, it is higher in the other 3 metrics. The sentence-level attention can improve the performance, demonstrating that it can capture more important posts.

EBAtt extracts semantic information on all posts by BiLSTM and fuses it with emotional representation. Results show that the F_1-score of EBAtt decreases by 2.9% compared with the proposed model, which indicates the necessity of selecting depression indicator posts.

RLAtt is the model after removing the emotion extraction network from ERAN. Similarly, the state of the RL selection layer does not contain the emotion vector. The F_1-score of RLAtt is lower than the proposed model by 3.1%, which indicates that the emotional information improves our model the most.

From the results, we can conclude that extracting emotional information through the pretrained TextCNN is beneficial for depression detection task. Selecting depression indicator posts based on emotional states is also necessary for depression detection. In addition, the sentence-level attention layer can focus on useful posts.

Figure 3. Results of ablation experiments. Emotion-Based Reinforcement Attention Network (ERAN) is the proposed model, and the remaining three are the models after removing one module of ERAN. Acc: accuracy; EBatt: emotion-based BiLSTM (bidirectional long short-term memory network) attention network; ERN: emotion-based reinforcement learning network; F1: F_1-score; P: precision; R: recall; RLAtt: reinforcement learning attention;



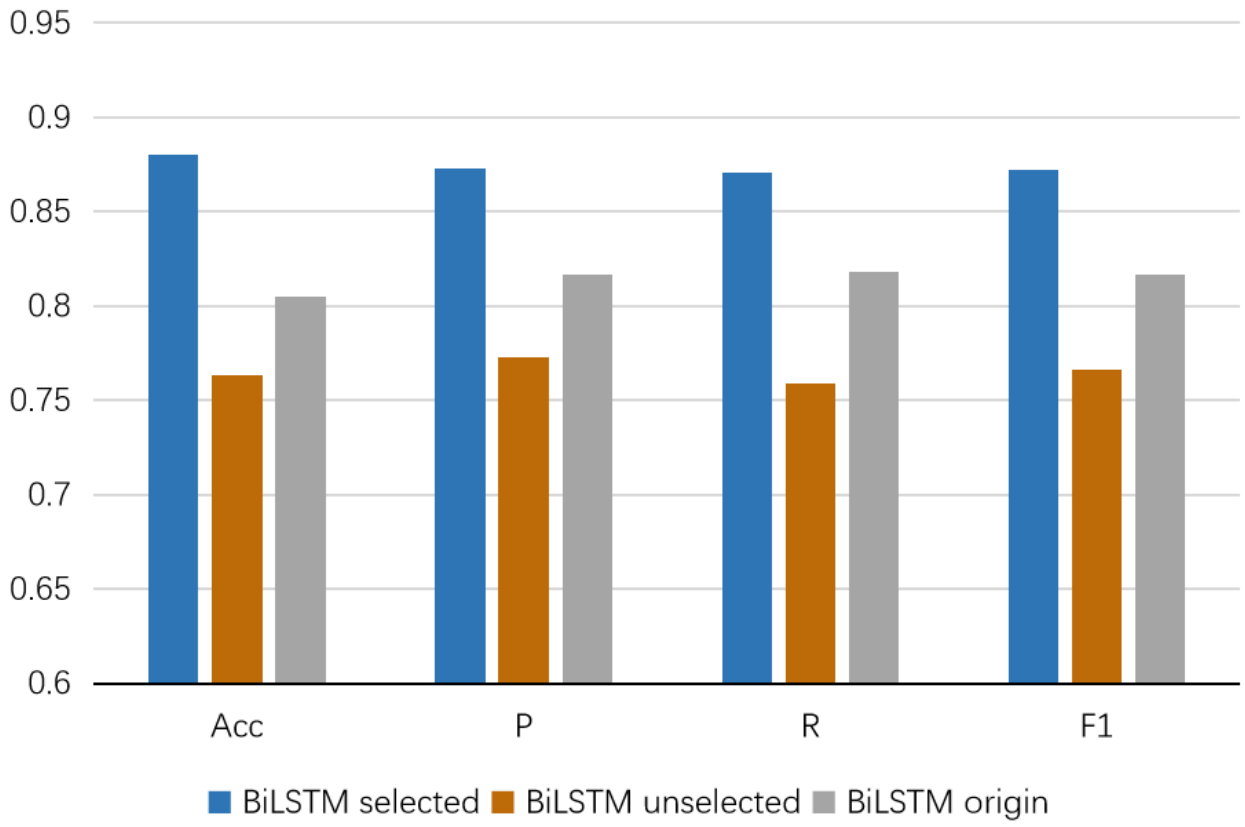
The Effectiveness of The RL Selection Layer

We train the proposed model to generate 2 subsets of depression-related and unselected posts from the original data set. Following this, we obtain 3 data sets, the selected indicator data set H^{dep}, the unselected data set H^{non}, and the original data set H^{orig}. The baseline model BiLSTM is then trained on each of these 3 data sets to verify the effectiveness of the RL selection layer. Figure 4 illustrates the results of the baseline model BiLSTM on the 3 data sets.

From Figure 4, we can conclude that the model trained on H^{dep} performs best. Meanwhile, the model trained on H^{non} achieves worse performance than the one trained on H^{orig}, which demonstrates the effectiveness of the RL selection.

To verify the effectiveness of introducing sentiment vectors in the RL selection module, we removed the sentiment vector in the state s^t. The ablation experiment achieves 88.3%, 88.1%, 87.3%, and 87.7% in accuracy, precision, recall, and F_1-scores, respectively. Through the results of the ablation experiment, we can find that the performance of the model decreases after removing the sentiment vectors from the RL selection module, which proves that the sentiment information is helpful for selecting depression indicator posts.

Figure 4. Comparative results of BiLSTM trained on the selected posts, the unselected posts, and the original posts. Acc: accuracy; BiLSTM: bidirectional long short-term memory network; F1: F_1-score; P: precision; R: recall.



Attention Visualization and Error Analysis

In this section, we extracted attention weights and visualized them to verify the validity of the sentence-level attention layer and the reasonableness of the selected posts. We have selected a part of the results of the users as examples, who are called "___mandyy" and "Adri." The results of attention visualization are illustrated in Figure 5.

The first example shows that the first post has the highest weight, where "my depression" indicates that the user has depression. The second post also contains the words "depression", "me", etc. Thus, "___mandyy" is finally classified as having "depression." As we can see, many of the selected posts of this user with depression are of negative sentiment, suggesting a strong association between depression and negative emotions.

The second user is the one we have used as an example in Figure 1. From the results of the visualization, we can observe that the fifth post has the highest weight. Classification results indicate that the user is indeed depressed. However, the posts "The view's really nice from here." and "I'm so proud of bts they deserve everything" are irrelevant to depression. In addition, the model assigns high weight to the first irrelevant post. One possible reason for choosing these posts is that they contain strong emotional expressions. We think it can be improved by developing a stricter selection strategy.

Figure 5. Examples of attention visualization. Different colors represent different weights. The deeper the color, the greater the weight of the post.

Discussion

Principal Findings

Based on the results, we can observe that introducing emotional information can be very helpful for depression detection tasks, indicating that emotional characteristics are strongly associated with depression. The strategy of selecting depression indicator posts from historical posts is critical to our model because it excludes the effect of irrelevant information. As only user-level labels are in the data set, we use RL to select posts rather than supervised learning. Furthermore, the fusion of emotion vectors into agent states is interpretable. The sentence-level attention layer assigns greater weight to relevant posts, which makes the model perform better.

Although the RL selection layer performs well, the selected posts still contain irrelevant posts with strong emotional expressions. Compared with other optimization methods, the convergence of policy gradient is better. However, this method tends to fall into local optimum and its training speed is slow.

Conclusions

In this paper, we addressed the task of depression detection of users on social media by proposing an ERAN. The proposed model contains 2 modules: the emotion extraction network and the RL attention network. It uses the pretrained word2vec embeddings as input. The emotion extraction network captures deep emotional information by a pretrained TextCNN. The RL attention network is composed of the BiLSTM layer, the RL selection layer, and the sentence-level attention layer. The RL selection layer can select depression indicator posts from original posts based on the emotional states, and the attention layer is able to assign greater weight to relevant posts. Results show that the proposed model outperforms the state-of-the-art model. We verified the validity of the emotion extraction network, the RL selection layer, and the sentence-level attention layer through an ablation study and a visualization analysis. The emotional features and selection of indicator posts are necessary for depression detection task.

The proposed model uses social media data set to detect depression, which can provide a certain degree of diagnostic basis and address the problem of the lack of effective objective diagnosis in the field of depression. In the future work, we will introduce users' personality information and multimodal information such as visual information to our model. We will further extract more detailed information about depression based on the proposed model to help analyze the pathogenesis of depression as well as accurate treatment.

Acknowledgments

The publication of this paper is funded by grants from the Natural Science Foundation of China (No. 62006034), Natural Science Foundation of Liaoning Province (No. 2021-BS-067), the Fundamental Research Funds for the Central Universities [No. DUT21RC(3)015], and the major science and technology projects of Yunnan Province (202002ab080001-1).

Authors' Contributions

BC performed the experiments and wrote the paper. JW and YZ provided theoretical guidance and the revision of this paper. HL, LY, and BX contributed to the algorithm design.

Conflicts of Interest

None declared.

References

1. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf [accessed 2022-07-13]
2. Yadollahpour A, Nasrollahi H. Quantitative Electroencephalography for Objective and Differential Diagnosis of Depression: A Comprehensive Review. GJHS 2016 Mar 31;8(11):249-256. [doi: 10.5539/gjhs.v8n11p249]
3. Digital 2022: global overview report. DataReportal. 2022. URL: https://datareportal.com/reports/digital-2022-global-overview-report [accessed 2022-07-13]
4. Park M, Cha C, Cha M. Depressive moods of users portrayed in Twitter. In: KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY: ACM; 2012 Presented at: KDD '12: The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 12-16, 2012; Beijing, China p. 12-16.
5. Choudhury DM, Counts S, Horvitz E. Predicting postpartum changes in emotion and behavior via social media. In: CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY: ACM; 2013 Apr Presented at: CHI '13: CHI Conference on Human Factors in Computing Systems; April 27, 2013 to May 2, 2013; Paris, France p. 3267-3276. [doi: 10.1145/2470654.2466447]
6. Nguyen T, Phung D, Dao B, Venkatesh S, Berk M. Affective and Content Analysis of Online Depression Communities. IEEE Trans. Affective Comput 2014 Jul 1;5(3):217-226 [FREE Full text] [doi: 10.1109/taffc.2014.2315623]
7. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, et al. Depression detection via harvesting social media: a multimodal dictionary learning solution. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Palo Alto,

CA: AAAI Press; 2017 Presented at: Twenty-Sixth International Joint Conference on Artificial Intelligence; August 19-25, 2017; Melbourne, VIC, Australia p. 3838-3834. [doi: 10.24963/ijcai.2017/536]
8. Tong L, Liu Z, Jiang Z, Zhou F, Chen L, Lyu J, et al. Cost-sensitive Boosting Pruning Trees for depression detection on Twitter. IEEE Trans. Affective Comput 2022. [doi: 10.1109/taffc.2022.3145634]
9. Park M, McDonald D, Cha M. Perception differences between the depressednon-depressed users in Twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. 2013 Presented at: Seventh International AAAI Conference on Weblogs and Social Media (ICWSM-13); July 8-11, 2013; Cambridge, MA p. 476-485 URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14425/14274
10. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics; 2017 Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; September 7-11, 2017; Copenhagen, Denmark p. 2968-2978. [doi: 10.18653/v1/d17-1322]
11. Alhanai T, Ghassemi M, Glass J. Detecting Depression with Audio/Text Sequence Modeling of Interviews. 2018 Presented at: Proceedings of the INTERSPEECH 2018; September 2-6, 2018; Hyderabad, Telangana, India p. 1716-1720 URL: https://groups.csail.mit.edu/sls/publications/2018/Alhanai_Interspeech-2018.pdf [doi: 10.21437/Interspeech.2018-2522]
12. Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S. Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation. JMIR Med Inform 2021 Jul 16;9(7):e28754 [FREE Full text] [doi: 10.2196/28754] [Medline: 34269683]
13. Orabi AH, Buddhitha P, Orabi MH. Deep learning for depression detection of twitter users. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. Stroudsburg, PA: Association for Computational Linguistics (ACL); 2018 Jun Presented at: Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; June 5, 2018; New Orleans, LA p. 88-97 URL: https://aclanthology.org/W18-06.pdf [doi: 10.18653/v1/W18-06]
14. Zogan H, Razzak I, Wang X, Jameel S, Xu G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web 2022;25(1):281-304 [FREE Full text] [doi: 10.1007/s11280-021-00992-2] [Medline: 35106059]
15. Gui T, Zhang Q, Zhu L, Zhou X, Peng M, Huang X. Depression Detection on Social Media with Reinforcement Learning. In: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings. Berlin/Heidelberg, Germany: Springer-Verlag; 2019 Oct Presented at: China National Conference on Chinese Computational Linguistics; October 18, 2019; Kunming, China p. 613-624. [doi: 10.1007/978-3-030-32381-3_49]
16. Kim Y. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics (ACL); 2014 Presented at: Conference on Empirical Methods in Natural Language Processing; 2014; Doha, Qatar p. 1746-1751 URL: https://aclanthology.org/D14-1181 [doi: 10.3115/v1/d14-1181]
17. Mikolov T, Sutskever I, Chen K. Distributed representations of words and phrases and their compositionality. In: NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Red Hook, NY: Curran Associates Inc; 2013 Presented at: 26th International Conference on Neural Information Processing Systems (NIPS'13); December 5-10, 2013; Lake Tahoe, NV p. 3111-3119.
18. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research 2011 Nov 1;12(2011):2493-2537 [FREE Full text] [doi: 10.5555/1953048.2078186]
19. Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). New York, NY: IEEE; 2013 Presented at: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding; December 8–13, 2013; Olomouc, Czech Republic p. 273-278 URL: https://doi.org/10.1109/ASRU.2013.6707742 [doi: 10.1109/asru.2013.6707742]
20. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Compution 1997 Nov 15;9(8):1735-1780 [FREE Full text] [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
21. Sutton RS, McAllester DA, Singh SP, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In: NIPS'99: Proceedings of the 12th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press; 1999 Nov Presented at: 12th International Conference on Neural Information Processing Systems (NIPS'99); November 29 to December 4, 1999; Denver, CO p. 1057-1063 URL: https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf [doi: 10.5555/3009657.3009806]
22. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics (ACL); 2016 Jun Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016); June 12-17, 2016; San Diego, CA p. 1480-1489 URL: https://aclanthology.org/N16-1174 [doi: 10.18653/v1/n16-1174]
23. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv. Preprint posted online December 22, 2014 [FREE Full text]

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 2011;12(2011):2825-2830 [FREE Full text]
25. Rolet A, Cuturi M, Peyré G. Fast dictionary learning with a smoothed Wasserstein loss. PMLR 2016;51:630-638 [FREE Full text] [doi: 10.1109/inmic.2016.7840071]
26. Song X, Nie L, Zhang L. Multiple social network learning and its application in volunteerism tendency prediction. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: Association for Computing Machinery; 2015 Presented at: SIGIR '15: The 38th International ACM SIGIR Conference on Research and Development in Information Retrieval; August 9-13, 2015; Santiago, Chile p. 9-13. [doi: 10.1145/2766462.2767726]
27. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. Preprint posted online July 26, 2019 [FREE Full text]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BiLSTM: bidirectional long short-term memory network
CNN: convolutional neural network
EBAtt: emotion-based BiLSTM attention network
ERAN: emotion-based reinforcement attention network
ERN: emotion-based reinforcement learning network
LSTM: long short-term memory network
MDHAN: multimodal depression detection with hierarchical attention network
MDD: multimodal depression data set
MDL: Multimodal Depressive Dictionary Learning
MSNL: Multiple Social Networking Learning
NB: naïve Bayesian
RL: reinforcement learning
RLAtt: reinforcement learning attention
RNN: recurrent neural network
RoBERTa: Robustly Optimized BERT pre-training Approach
WDL: Wasserstein Dictionary Learning
WHO: World Health Organization

Edited by T Hao; submitted 17.03.22; peer-reviewed by J Gao, Y Du, M Torii; comments to author 05.06.22; revised version received 02.07.22; accepted 06.07.22; published 09.08.22.

Please cite as:
Cui B, Wang J, Lin H, Zhang Y, Yang L, Xu B
Emotion-Based Reinforcement Attention Network for Depression Detection on Social Media: Algorithm Development and Validation
JMIR Med Inform 2022;10(8):e37818
URL: https://medinform.jmir.org/2022/8/e37818
doi:10.2196/37818
PMID:35943770

Original Paper

Identifying Patients Who Meet Criteria for Genetic Testing of Hereditary Cancers Based on Structured and Unstructured Family Health History Data in the Electronic Health Record: Natural Language Processing Approach

Jianlin Shi[1,2,3], MS, MD, PhD; Keaton L Morgan[3,4], MS, MD; Richard L Bradshaw[3], MS, PhD; Se-Hee Jung[3,5], BSN; Wendy Kohlmann[6,7], MS; Kimberly A Kaphingst[7,8], SCD; Kensaku Kawamoto[3], MPH, MD, PhD; Guilherme Del Fiol[3], MD, PhD

[1]Veterans Affairs Informatics and Computing Infrastructure, Department of Veterans Affairs Salt Lake City Health Care System, Salt Lake City, UT, United States

[2]Division of Epidemiology, Department of Internal Medicine, School of Medicine, University of Utah, Salt Lake City, UT, United States

[3]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

[4]Department of Emergency Medicine, University of Utah, Salt Lake City, UT, United States

[5]College of Nursing, University of Utah, Salt Lake City, UT, United States

[6]Department of Population Health Sciences, University of Utah, Salt Lake City, UT, United States

[7]Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[8]Department of Communication, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Guilherme Del Fiol, MD, PhD
Department of Biomedical Informatics
University of Utah
421 Wakara Way
Ste 140
Salt Lake City, UT, 84108-3514
United States
Phone: 1 801 581 4080
Fax: 1 801 581 4297
Email: guilherme.delfiol@utah.edu

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2022/9/e42533

Abstract

Background: Family health history has been recognized as an essential factor for cancer risk assessment and is an integral part of many cancer screening guidelines, including genetic testing for personalized clinical management strategies. However, manually identifying eligible candidates for genetic testing is labor intensive.

Objective: The aim of this study was to develop a natural language processing (NLP) pipeline and assess its contribution to identifying patients who meet genetic testing criteria for hereditary cancers based on family health history data in the electronic health record (EHR). We compared an algorithm that uses structured data alone with structured data augmented using NLP.

Methods: Algorithms were developed based on the National Comprehensive Cancer Network (NCCN) guidelines for genetic testing for hereditary breast, ovarian, pancreatic, and colorectal cancers. The NLP-augmented algorithm uses both structured family health history data and the associated unstructured free-text comments. The algorithms were compared with a reference standard of 100 patients with a family health history in the EHR.

Results: Regarding identifying the reference standard patients meeting the NCCN criteria, the NLP-augmented algorithm compared with the structured data algorithm yielded a significantly higher recall of 0.95 (95% CI 0.9-0.99) versus 0.29 (95% CI

0.19-0.40) and a precision of 0.99 (95% CI 0.96-1.00) versus 0.81 (95% CI 0.65-0.95). On the whole data set, the NLP-augmented algorithm extracted 33.6% more entities, resulting in 53.8% more patients meeting the NCCN criteria.

Conclusions: Compared with the structured data algorithm, the NLP-augmented algorithm based on both structured and unstructured family health history data in the EHR increased the number of patients identified as meeting the NCCN criteria for genetic testing for hereditary breast or ovarian and colorectal cancers.

(JMIR Med Inform 2022;10(8):e37842) doi:10.2196/37842

KEYWORDS

clinical natural language processing; family health history extraction; cohort identification; genetic testing of hereditary cancers

Introduction

Background

Cancer screening has been shown to effectively reduce mortality [1,2]. Unlike population-based screening recommendations that target a broad range of individuals, increasing evidence supports individualized cancer screening according to cancer risk [3-5]. Individuals at higher risk may benefit from earlier, more frequent, or more intensive screening. Effective interventions are needed to stratify patients by risk and to direct them to an appropriate level of screening. However, individualizing screening on a population scale requires patient-specific risk assessments for several types of cancer. This is quite challenging in today's overwhelmed primary care environment, as the current screening process requires manual chart review to identify patient candidates for genetic testing, and primary care providers often do not have time or knowledge to discuss genetic testing with their patients. A promising solution is to automate the identification of high-risk patients using electronic health records (EHRs) coupled with clinical decision support (CDS) tools.

The National Comprehensive Cancer Network (NCCN) has published a set of evidence-based guidelines for genetic testing of hereditary cancers, including breast, ovarian, pancreatic, and colorectal cancers [6,7]. A summary of these 2 guidelines is listed in Textbox 1, where each table cell represents a criterion, and the criteria for the same cancer cohort are listed in the same column. When one or more criteria are met, the corresponding genetic testing is recommended. These cancer risk assessment guidelines are based mainly on the family health history (FHH) of cancer or cancer syndromes, which is recorded in EHR systems as part of routine patient care activities. Therefore, EHR is one of the most important sources of FHH that can be used to drive CDS tools to help identify candidates for genetic testing of hereditary cancers [8]. However, several challenges limit the systematic use of FHH in EHR for these purposes, including (1) scattered FHH documentation in both structured and unstructured formats across different EHR sections, such as the clinical note [9], problem list, and FHH sections; (2) conflicting documentation in different sections of the EHR; (3) incomplete documentation in structured FHH data; (4) negation and ambiguity of information in unstructured data [10-12].

Textbox 1. Excerpt of National Comprehensive Cancer Network (NCCN) criteria for unaffected individuals' family history–based genetic testing of breast, ovarian, pancreatic, and colorectal cancers (referenced with permission).

Breast or ovarian cancer:

1. First- or second-degree relative with breast cancer at age ≤45 years
2. First- or second-degree relative with ovarian cancer
3. First-degree relative with pancreatic cancer
4. Breast cancer in a male relative
5. Three or more first- or second-degree relatives with breast or prostate cancer on the same side of the family
6. Ashkenazi Jewish and any breast or prostate cancer in any relative at any age
7. BRCA1/2, CHEK2, ATM, PALB2, TP53, PTEN, or CDH1 genes, Cowden Syndrome, Li-Fraumeni Syndrome in any relative at any age

Colorectal cancer:

1. MLH1, MSH2, PMS2, MSH6, EPCAM, MYH, or MUTYH genes, Lynch syndrome, familial adenomatous polyposis (FAP), adenomatous polyposis coli (APC), serrated polyposis or polyposis discovered in the coded family history
2. First-degree relative with colon cancer at ≤50 years
3. First-degree relative with endometrial cancer at ≤50 years
4. Three or more first- or second-degree relatives with Lynch syndrome, HNPCC, colon cancer, endometrial, uterine, ovarian, stomach, gastric, small bowel, small intestine, kidney, ureteral, bladder, urethra, brain, pancreas, also all on the same side of the family

Genetic testing for breast, ovarian, or colorectal cancer is recommended if at least one of these criteria is met.

Current EHR systems often provide a dedicated FHH section, in which FHH assertions can be captured using a combination

of structured (eg, coded disease, relationship, and age of onset) and unstructured data (ie, the comments field). FHH free-text comments are different from broader clinical notes in that the former are associated with a specific structured FHH assertion, only available in the FHH section, while clinical notes can capture a much wider range of information, including medical history, physical examination, and treatment plans. Health care providers typically use free-text FHH comment fields when desired information cannot be fully captured as structured data. For example, a patient's sister who developed breast cancer in her 30s can be captured partially as structured data (ie, condition = *breast cancer* and family member = *sister*) supplemented by a comment captured in the unstructured data conveying the uncertain age of onset (ie, *onset in her 30s*). The FHH section is increasingly used as part of routine visit intake by medical assistants and by patients themselves through patient portals [13]. Therefore, the FHH section is a promising and underused source of FHH for EHR.

Previous studies have largely focused on extracting FHH from clinical notes [14,15]. This study is the first comprehensive attempt to supplement structured FHH data with information extracted from free-text comments. The natural language processing (NLP) extraction of information from free-text comments imposes a unique set of challenges that require specific approaches that have not been investigated. Specifically, candidate approaches must address the interplay between structured and unstructured data collected in the FHH section.

Objectives

Our previously developed structured algorithm [8] for identifying patients who met the NCCN criteria for genetic testing using structured data demonstrated the potential use of this dedicated FHH section. Nonetheless, we noticed that the algorithm based on structured data failed to correctly identify certain cases because some information needed for eligibility determination was recorded as free-text comments. For example, an FHH entry included *CANCER* and *AUNT* as structured data, with the specific type of cancer and age of onset (*breast ca, dx in 30s*) provided as a free-text comment. This case would be considered eligible for genetic testing when using the information provided in the comments section. These errors resulting from the structured data algorithm added a manual review burden for genetic counseling staff because they needed to manually confirm patient eligibility before communicating with them.

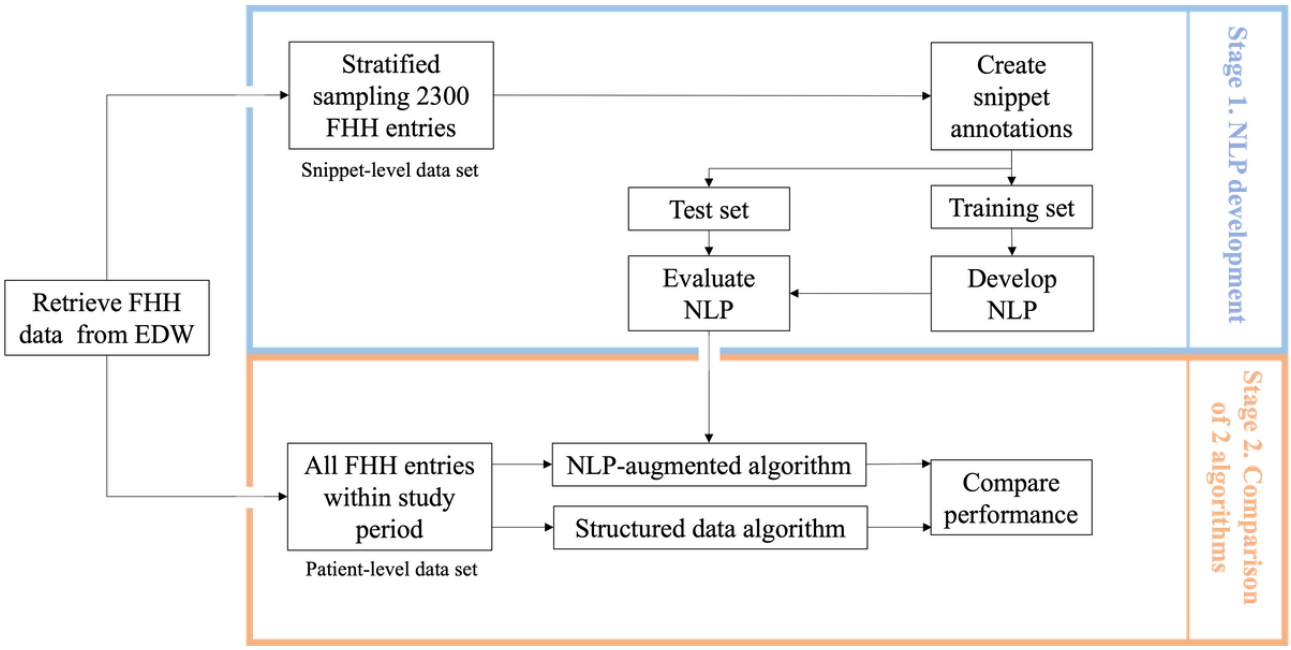
Hence, this study aims to augment CDS algorithms that rely exclusively on structured FHH data with information extracted from free-text FHH comments fields using NLP, with a focus on identifying patients who meet the NCCN criteria for genetic testing for hereditary breast or ovarian and colorectal cancers. The corresponding NLP was designed to extract the FHH information when it was not available or accurately coded in structured data, including the cancer type (eg, pancreatic cancer), the age of onset (eg, in the early 30s), and the affected family member (eg, *paternal aunt*). The primary hypothesis is that using NLP to augment the previously developed algorithm (using structured data alone) [8] can improve the accuracy of identifying patients who meet the NCCN criteria for genetic testing based on the FHH of patients seen in primary care settings at a US academic medical center.

Methods

Study Design

We retrospectively studied data from the EHR at the University of Utah Health. The study consisted of 2 stages (Figure 1). In the first stage, for NLP development, an NLP solution was developed to extract FHH information from both structured and unstructured data in the FHH section of EHR, and its performance was evaluated in comparison with gold standard annotation results. Next, we developed an NLP-augmented algorithm on top of the structured data algorithm (using only structured data) [8] to match the NCCN criteria using the NLP-processed results from both structured and unstructured fields. In the second stage, the performance of the NLP-augmented algorithm was compared with that of the structured data algorithm.

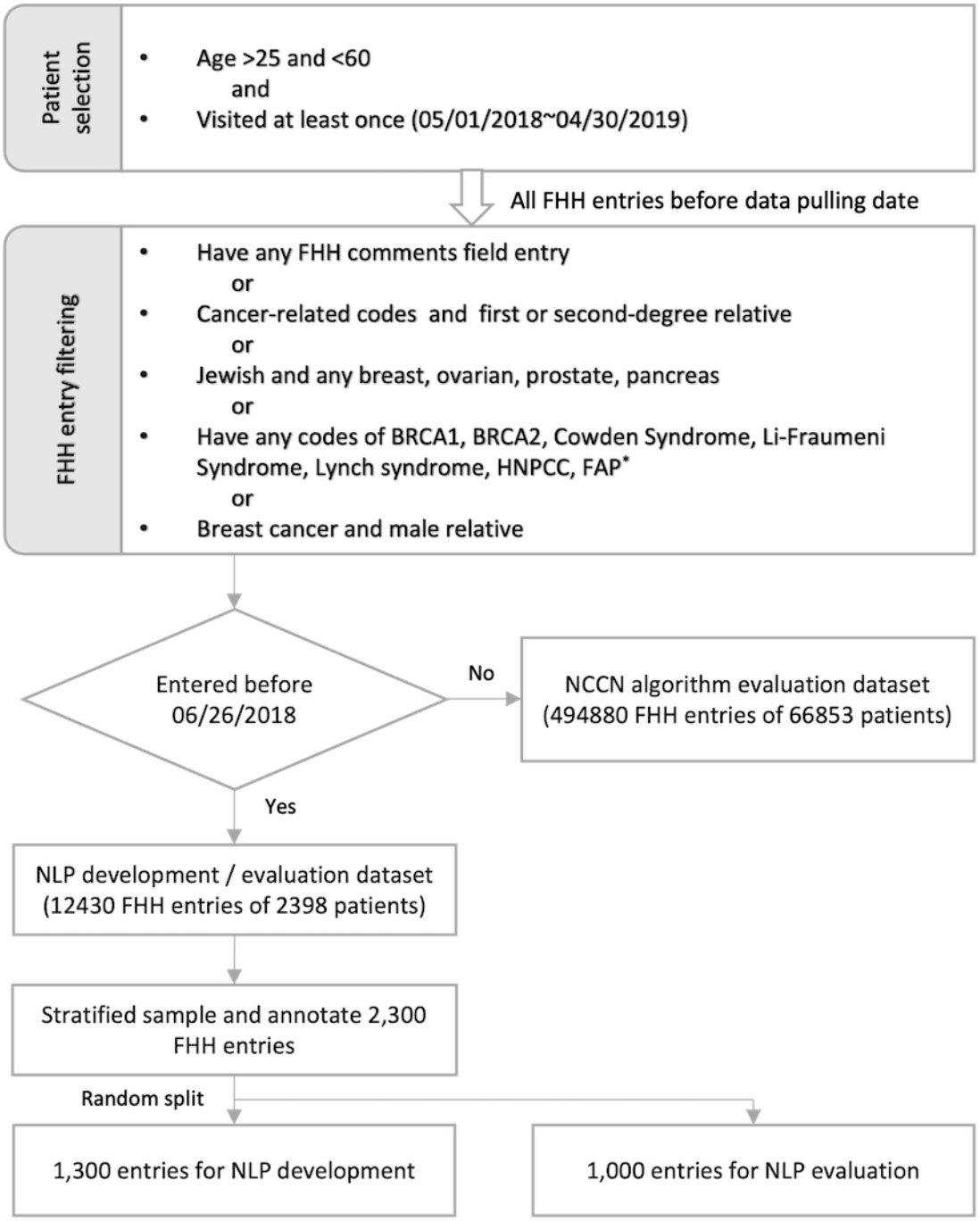
Figure 1. Study stages, including natural language processing (NLP) development (stage 1) and comparison between the NLP-augmented algorithm and an algorithm using only structured data (stage 2). EDW: enterprise data warehouse; FHH: family health history.



Data Sets

The data set for NLP development and evaluation consisted of EHR-based data from the FHH section (including both structured and unstructured fields) for 77,423 patients aged between 25 and 60 years who visited the University of Utah Health primary care clinic at least once between May 1, 2018, and April 30, 2019. All FHH entries of these patients were obtained, including entries recorded in prior visits to June 26, 2014. FHH entries contained a coded condition (breast cancer), a coded relative (sister), age of onset integer, and a free-text comment clinicians used to add detail (*in her 30s*). Entries that were not used to determine familial cancer risk were filtered using Structured Query Language (SQL), resulting in 31,191 entries. The detailed filtering strategy is illustrated in Figure 2.

Figure 2. Data set creation process. FHH: family health history. NCCN: National Comprehensive Cancer Network. NLP: natural language processing. *HNPCC: hereditary non-polyposis colorectal cancer. FAP: familial adenomatous polyposis. Other genetic mutations or cancer syndromes specified in the NCCN guideline but without a code in electronic health record (EHR) were not included.



The data set was split into 2. The FHH entries that were entered before June 26, 2018 were used for NLP development and evaluation (ie, the NLP development or evaluation data set), while entries entered after that date were used for algorithm evaluation (ie, the NCCN algorithm evaluation data set). We obtained a stratified random sample of 2300 FHH entries from the NLP development and evaluation data set. The stratification was based on the diagnosis codes in the condition field and stratified into four groups: (1) breast or ovarian cancer, (2) colorectal cancer, (3) other cancers, and (4) other noncancer family histories, at a 1:1:2:2 ratio. We randomly split 1300 FHH entries for NLP development, and the remaining 1000 entries were used for the snippet-level NLP evaluation. The NCCN algorithm evaluation data set was used to compare the performance of the 2 algorithms. Then, all the FHH entries (both data sets) were used to estimate the amount of additional information extracted by NLP and compare the patients identified by the NLP-augmented algorithm with those identified by the structured data algorithm.

NLP Approach

Overview

Although NLP is often only used to process free-text data, independent of structured data, the comments field in the FHH

section of EHR is used to supplement the structured data and cannot be interpreted in isolation. For example, in Table 1, the word *breast* supplements the concept *CANCER* in the structured condition field. Therefore, we concatenated the structure and comments fields into a single string for NLP processing. We also used double curly brackets to mark the values from the structured fields to reconcile conflicting information between the structured and comments fields (Table 1).

Table 1. An example of combining structured and unstructured data from FHH[a] assertions.

Field names	Condition	Comments[b]	Family member	Age of onset
Original data	CANCER	Breast, great-aunt, dx at age of 52	AUNT	NULL
Combined	{{CANCER}}	Breast, great-aunt, dx at age of 52	{{AUNT}}	{{}}
Annotations				

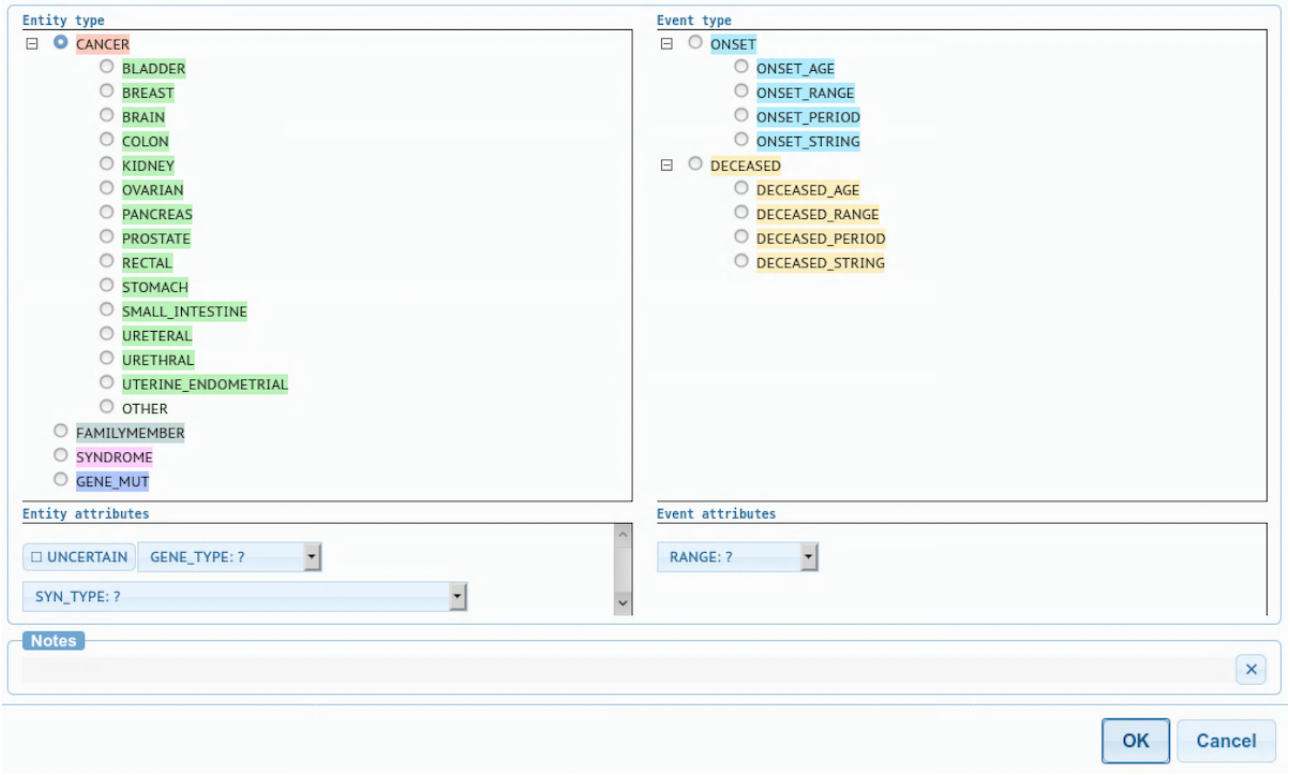
[a]FHH: family health history.

[b]In this case, the comments field supplements or corrects the structured data, that is, CANCER is of the *breast*, and the family member (AUNT) is actually the patient's great-aunt. FX_CANCER (FC): family member to cancer relationship; FX_ONSET: family member to age of onset relationship.

FHH Annotation Schema

A total of 2 physicians designed the annotation schema based on the FHH attributes relevant to the NCCN guidelines for genetic testing of hereditary breast or ovarian and colorectal cancers. This schema encompasses conditions, family members, and the age of onset. Specifically, the snippet-level data set contains (1) annotated entities for cancer diagnosis (CANCER), cancer-related syndromes (SYNDROME), cancer-related genetic mutations (GENE_MUT), family members (FAMILYMEMBER), and age of onset (ONSET), and (2) relations between family members and conditions, as well as between family members and age of onset. The example provided in Table 1 has 3 entities, that is, *great-aunt* (FAMILYMEMBER), ([CANCER]) *breast* (BREAST−breast cancer), *52* (ONSET_AGE), and 2 relations, that is, *great aunt → {{CANCER}} breast* (FX_CANCER) and *great aunt → 52* (FX_ONSET). As the NCCN criteria include other cancers with mutations that share a common genetic pathway with breast, ovarian, and colorectal cancers, we added the following annotation subtypes: BLADDER, BREAST, BRAIN, COLON, KIDNEY, OVARIAN, PANCREAS, PROSTATE, RECTAL, STOMACH, SMALL_INTESTINE, URETERAL, and URETHRAL. As the NCCN criteria also use the side of the family of the affected family member and the degree of relationship, 2 attributes were included: family member CODE (eg, *GRANDMOTHER*) and SIDE of FAMILYMEMBER (eg, *PATERNAL*). In addition, an UNCERTAINTY feature was added to capture uncertainty statements (eg, *probably ovarian cancer)*. We used a schema developed in our previous studies to annotate the age of onset [10], which includes 4 subtypes: ONSET_AGE (eg, *age 52*), ONSET_RANGE (eg, *in his 30s*), ONSET_PERIOD (eg, *in 1965*), and ONSET_STRING (eg, *postmenstruation*). Figure 3 presents a screenshot of the full schema within the annotation tool (Brat) [16]. The schema configuration is shared in GitHub [17].

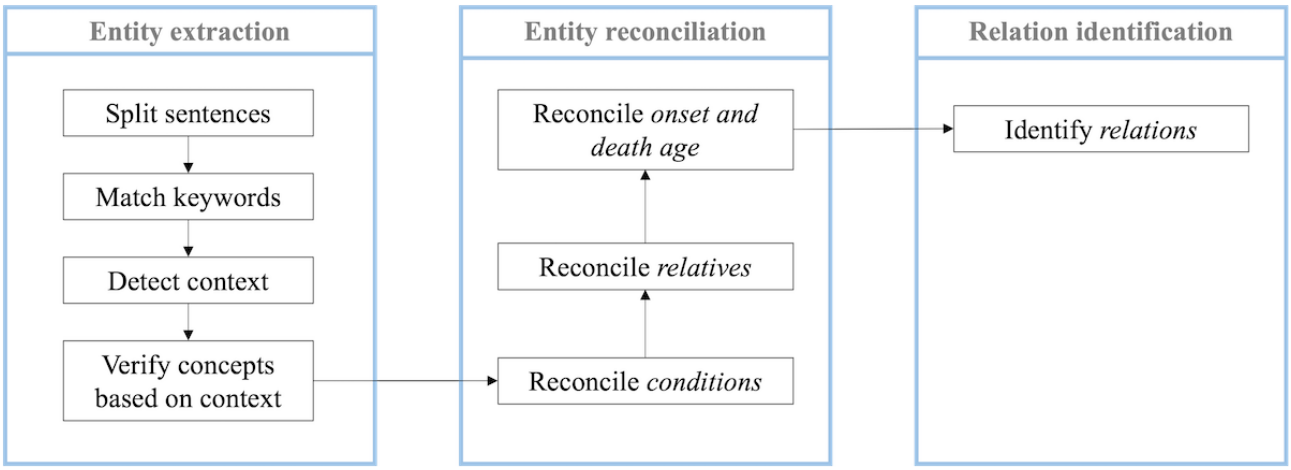
Figure 3. Screenshot of the schema as implemented with the annotation tool Brat.



NLP Development

To develop the NLP pipeline, we used Easy clinical information extractor (EasyCIE), a lightweight rule-based NLP tool that supports rapid clinical NLP implementations [18]. All NLP components of EasyCIE are configurable through rules without the need to develop new pipelines. A total of 1300 FHH entries were used to develop the rules. We adopted a logic similar to that described by Goryachev et al [19] but implemented the logic in a different way for efficiency and generalizability considerations [20]. The processing consists of three major steps: (1) entity extraction, (2) entity reconciliation, and (3) relation identification (Figure 4). Each step was performed using one or more NLP components. The following paragraph explains these components using the examples in Table 1. Each component is configured using a separate rule set that incorporates a keyword dictionary or inference logic. These rules were developed based on 3 sources: Unified Medical Language System, training data set, and clinical domain experts' input. The rule set is available on GitHub [21].

Figure 4. Easy clinical information extractor processing workflow. Three major steps (blue boxes): (1) entity extraction—extract the entities from the family health history entries; (2) entity reconciliation—reconcile the conflicts between the extracted entities; (3) relation identification—link related entities. In each step, there are ≥1 natural language processing components to complete processing substeps.



Entity extraction (step 1) extracts the key entities (5 types) from the FHH entries. First, we split the sentences if there were more than one sentence. Second, we attempted to match the input string with controlled vocabulary (a keyword dictionary). An example is shown in Table 1, *{{CANCER}}* breast was recognized as *BREAST* (cancer), *52* as *ONSET_AGE*, and *great aunt* and *AUNT* as *FAMILYMEMBER*. Next, we detected the double curly brackets around *AUNT*. These 2 symbols indicate the mention of *AUNT* was located in the structured field. Thus, we assigned the feature *is_structural* to *AUNT*. Finally, we

verified the features of each entity to determine whether they matched any inference rules. In the example, a *FAMILYMEMBER* with the *is_structural* feature was classified as *STR_FAMILYMEMBER* (a family member in the structured field). This differentiation among entities in different contexts allows entity reconciliation in the next step. This component also allowed us to exclude irrelevant mentions of entities (eg, the *FAMILYMEMBER daughter* in the context of *live with her daughter*). Further details regarding the implementation of EasyCIE's rule-processing engine are available elsewhere [20].

Entity reconciliation (step 2) reconciles the extracted entities from the first step when conflicts exist between structured and unstructured data. The corresponding heuristic rules were iteratively developed based on annotated data from the training set with refinements based on error analysis after applying the algorithm to the training set. In addition, we obtained insights through discussions with clinical experts, who currently use the dedicated FHH section to document FHH. Specifically, the following (Textbox 2) heuristic rules were applied (Table 2, additional examples are listed).

Textbox 2. Heuristic rules.

Rules

1. If the structured field indicated *colon cancer*, but the information in the comments field clarified the condition of interest to be colorectal cancer syndromes (eg, Lynch syndrome), SYNDROME overrode COLON (cancer) in the structured field
2. If the age of onset was documented as structured numeric data (eg, 50) but the comments field documented an *ONSET_RANGE* (eg, late 50s), the ONSET_RANGE overrode the structured age of onset
3. If the age of onset was available in structured data, and the comments field included *ONSET_PERIOD* (eg, in 1985) or *ONSET_STRING* (eg, 10 years ago), *ONSET_PERIOD* and *ONSET_STRING* >were ignored. (4) If no age of onset was documented in the structured field and the comments field included a *DECEASED_AGE*, the algorithm set an *ONSET_RANGE before* the *DECEASED_AGE*.
4. If the comments field contained information on a specific family member, the algorithm ignored the structured family member field unless the comments field included a conjunction such as *also* >or *and*. In the example in Table 1, the *FAMILYMEMBER great aunt* was likely a correction of the STR_FAMILYMEMBER *AUNT* >because the picklist associated with STR_FAMILYMEMBER did not include an option for *great aunt*. Thus, in the reconciliation, STR_FAMILYMEMBER *AUNT* is ignored.
5. If the comments field contained nonspecific family member information (eg, father's side), whereas the structured field contained a specific family member, the structured field code was used, and information from the comments was added as attributes if applicable.
6. If a mention of *FAMILYMEMBER* was specified as multiple individuals (eg, 2× sisters), multiple instances of *FAMILYMEMBER* were created (eg, 2× sisters would lead to 2 instances).

Table 2. Heuristic rules to reconcile entities.

Structured fields	Example	Comments field	Example	Reconciliation
COLON[a] (cancer)	{{CANCER, COLON}}	Colorectal cancer–related *SYNDROME*	Lynch syndrome	Chose *SYNDROME*
ONSET_AGE	{{50}}	*ONSET_RANGE*	The late 50s	Chose *ONSET_RANGE*
ONSET_AGE	{{50}}	*ONSET_PERIOD*	In 1985	Chose *ONSET_AGE*
ONSET_AGE	{{50}}	*ONSET_STRING*	10 years ago	Chose *ONSET_AGE*
NULL	{{}}	*DECEASED_AGE*	Deceased at age 60 years	Inferred the *ONSET_RANGE*
FAMILYMEMBER	{{AUNT}}	A specific *FAMILYMEMBER*	Great-aunt	Chose *FAMILYMEMBER* in comments
FAMILYMEMBER	{{MOTHER}}	A specific *FAMILYMEMBER* with conjunction statement	And grandmother	Use *FAMILYMEMBER* in both fields
FAMILYMEMBER	{{AUNT}}	Nonspecific	Father side	Chose *FAMILYMEMBER*, and added comments value as a feature, if applicable
NULL	{{}}	Multiple *FAMILYMEMBER*	2× sisters	Created two *FAMILYMEMBER* annotations

[a]Words in italics denote concepts in the NLP output according to the FHH annotation schema.

Relation identification (step 3) links related entities. In the example of Table 1, *great aunt* and *{{CANCER}} breast* were linked to create an *FX_CANCER* relation. It also linked *great aunt* and *52* to create an *FX_ONSET* relation. As *STR_FAMILYMEMBER AUNT* was changed to *IGN_FAMILYMEMBER*, *AUNT* in the structured field were not linked to *{{CANCER}} breast* or *52*. When counting the number of relatives of interest, the number of FAMILYMEMBER-CANCER relations was obtained instead of relative entities. For example, *ovarian and stomach cancer* in *grandmother* should be counted as 2 cancers in the NCCN criteria. Although the NLP algorithm generated one

FAMILYMEMBER entity (grandmother), two FAMILYMEMBER-CANCER relations were generated. The same process is followed to handle cases where a single cancer assertion refers to multiple relatives, eg, *breast cancer in mother and aunt*.

NLP Performance Evaluation

We evaluated the NLP solution by comparing its output with the test set annotations of the snippet-level data set (1000 FHH entries). To save time and effort, entities with no relation were not annotated (eg, an entry that only has a condition without mentioning any family member); therefore, we did not evaluate the NLP performance for named entity recognition. Precision, recall, and F1 scores were calculated for relation identification. A true positive relation was counted when NLP-extracted information matched the reference standard for both the relation type and corresponding feature values, as well as the two linked entities. We applied the bootstrap sampling method [22] to estimate the 95% CI for each performance measurement and conducted error analyses by categorizing and counting different types of errors. Considering the mentions of *SYNDROME* (cancer syndrome) and *GENETIC_MUTATION* (cancer-related genetic mutation) were very rare in the data set, the CI for the performance related to the extraction of relations with these 2 entity types, that is, *FX_SYNDROME* (family member to cancer syndrome relation) and *FX_GENE_MUT* (family member to genetic mutation relation), could not be obtained. Thus, we only calculated the CIs of the microaverages of these 3 measurements using bootstrap methods over the aggregated data that included all 4 relation types.

Structured Data Algorithm for Patient Eligibility Assessment

A rule-based algorithm was previously developed [8] based on NCCN guidelines for the genetic testing of hereditary breast or ovarian and colorectal cancers [6,7] using only structured FHH data. The algorithm was implemented using an open-source CDS platform (OpenCDS [23]) through a standards-based approach based on CDS Hooks for Services and the Fast Healthcare Interoperability Resources standard for FHH data representation. On the basis of the patient's age and FHH, the algorithm determines whether the patient meets the NCCN criteria for genetic testing. The algorithm has been deployed for clinical use and integrated with the Epic EHR at the University of Utah Health and New York University. The details of the algorithm and its deployment in clinical practice are available elsewhere [8,24]. In this study, we used a structured data algorithm as the baseline.

NLP-Augmented Algorithm

The NLP-augmented algorithm was built on top of the structured data algorithm by converting the NLP output into a structured FHH format (condition, family member, and age of onset). As a result, the same structured data algorithm consumes NLP-augmented data. To handle the uncertainties, 2 different NLP configurations were provided, including and excluding uncertainty assertions for each of the breast, ovarian, and colorectal cancer cohorts. The configuration that included cases with uncertainty assertions was used to estimate the impact of NLP augmentations on algorithm-identified genetic testing candidates.

NLP-Augmented Algorithm Evaluation

The evaluation of the NLP-augmented algorithm consisted of two parts: (1) comparing the performance of the NLP-augmented algorithm with that versus the structured data algorithm using manually reviewed data as a reference standard, and (2) estimating NLP's impact on the patient cohort size generated by the structured data algorithm over the whole data set using the inclusion configuration. A patient-level data set was created in this study. Owing to the large size of the cohort, it was not feasible to validate the expected output for all patient cases. Therefore, we sampled and annotated the algorithm outputs (against the NCCN algorithm evaluation data set) instead of annotating the input data. A review of a subset of 200 cases showed that when the baseline and NLP-augmented algorithms agreed regarding patient eligibility for genetic testing, the algorithm outputs were correct in 100% of the cases. Therefore, for cost-efficient considerations, we applied stratified sampling to down-sample the cases in which the 2 algorithms agreed to maintain a 1:2 ratio between cases with agreement and disagreement. We sampled 100 cases in total, 50 breast and ovarian cancer screening candidates and 50 colorectal cancer screening candidates. Subsequently, 2 annotators independently reviewed these cases to determine whether the 2 algorithms reached the correct conclusion. Any disagreement between the 2 annotators was adjudicated by a third annotator.

The structured data and NLP-augmented algorithms were compared in terms of precision, recall, and F1 scores. The 95% CIs were computed using the bootstrap method. As we did not obtain the ground truth of the patients' FHH by contacting the patients themselves, the reference standards were made solely based on the entries in the FHH section. Next, we estimated the effectiveness of NLP augmentation by comparing (1) the number of FHH entries that were computable for the NCCN criteria and (2) the number of patients who met the criteria with and without NLP.

Ethics Approval

This study was approved by the institutional review board at the University of Utah (IRB_00154076).

Results

Data Set Description

After splitting the data set, 2398 patients with 12,430 FHH entries were included in the NLP development or evaluation data set and 66,853 patients with 494,880 FHH entries were included in the NCCN algorithm evaluation data set. A total of 8172 patients did not have any FHH entries and were excluded from the data set. These 2 data sets were similar in sex, race, ethnicity, and age (Table 3).

Table 3. Patient characteristics in the NLP[a] development or evaluation data set and the NCCN[b] algorithm evaluation data set.

Characteristic	NLP development or evaluation data set (n=2398)	NCCN algorithm evaluation data set (n=66,853)
Gender (male), n (%)	998 (41.2)	24,524 (36.7)
Race, n (%)		
White	1752 (73.2)	51,171 (76.5)
Other	359 (15)	9510 (14.2)
Asian	141 (5.9)	2973 (4.4)
Black or African American	67 (2.8)	1450 (2.2)
Not reported	56 (2.3)	1226 (1.8)
American Indian or Alaska Native	17 (0.7)	523 (0.8)
Hispanic ethnicity	327 (13.6)	9147 (13.7)
Age (years), mean (SD)	40.2 (9.6)	42.6 (9.9)

[a]NLP: natural language processing.

[b]NCCN: National Comprehensive Cancer Network.

NLP Performance Evaluation Results

Using the snippet-level test data set, we evaluated the NLP's performance at the snippet level; the average precision was 0.94 with 95% CI 0.91-0.97, the average recall was 0.94 with 95% CI 0.90-0.96, the average F1 score was 0.94 with 95% CI 0.91-0.96. The performance of the measurements for each relationship type is presented in Table 4.

Table 4. The performance on the snippet-level data set.

Relation types	TP[a]	FP[b]	FN[c]	Precision	Recall	F1 score
FX_CANCER[d]	489	32	31	0.94	0.94	0.94
FX_SYNDROME[e]	2	1	3	0.67	0.40	0.50
FX_GENE_MUT[f]	2	0	0	1.00	1.00	1.00
FX_ONSET[g]	203	10	14	0.95	0.94	0.94
Microaverage[h]	N/A[i]	N/A	N/A	0.94 (0.91-0.97)	0.94 (0.90-0.96)	0.94 (0.91-0.96)

[a]TP: true positive.

[b]FP: false positive.

[c]FN: false negative.

[d]FX_CANCER: family member to cancer relation.

[e]FX_SYNDROME: family member to cancer syndrome relation.

[f]FX_GENE_MUT: family member to cancer-related gene-mutation relation.

[g]FX_ONSET: Family member to age of onset relationship.

[h]These scores were computed using aggregated data, including all 4 relation types. The CIs were computed using the bootstrap method.

[i]N/A: not applicable.

NLP Error Analysis

On the basis of the snippet-level error analysis of the NLP output from the test data set of 1000 FHH entries, we found 6 error types (Table 5). Approximately 50% of the errors were not directly caused by NLP mistakes. The *Annotation Error* was made by the annotators, which is common when a large volume of data needs to be reviewed. In addition, as we only partially overlapped the annotations and adjudicated the disagreement between the annotators for greater efficiency, the data that were not overlapped might also have contributed to annotation errors. *Data Input Typos* were another complication, especially some rare typos; for example, *bladdler*. *Out of Vocabulary* signified the words and phrases that were not seen in the training set and not added to the knowledge base from Unified Medical Language System and experts' suggestions. For instance, *precancer* in the entry of *{{CANCER, BREAST}} precancer, age 30 {{MOTHER}} {{}}* should override the breast cancer code, because *precancer* is a term that describes a lesion that may develop into cancer. The NLP did not recognize the term; therefore, it was not possible to exclude breast cancer as an existing family health history. A *Context Error* might happen when the context of the entities included subtleties that the NLP could not correctly parse, for example, *{{CANCER, COLON}} possible, colon cancer, died when pt was 5 years old {{FATHER}} {{}}*. The NLP did not expect that the *5-year old* was not describing the father's age of onset in the comments field, but the patient's age. Sometimes, the input data is so

ambiguous (ambiguous input) that even our annotators were not sure of the exact meaning without referring to other sources. For example, the entry *{{CANCER, COLON}} ileum {{FATHER}} {{}}*, likely meant the father had *ileum cancer*, which overwrote *colon cancer*. However, we were not 100% confident if the father actually had both because most of the cases like these would have been coded as *{{CANCER, OTHER}} ileum {{FATHER}} {{}}*. In real practice, genetic counselors would need to go over some clinical notes to find statements that can be cross-referenced or reach out to the patient to confirm the information. These types of improper coding in the structured fields and the conflicting information between the structured fields and comments field indicate that the EHR user interface for FHH entry may benefit from redesign, such as allowing users to label uncertainty. Finally, when designing the schema for annotation, we aimed to capture as much useful information as possible. We included three aggregated types of cancer, *GYNECOLOGIC*, *GASTROINTESTINAL*, *GENITOURINARY*, to code cancers not specific to the anatomical sites indicated in the guidelines. However, when executing the algorithms, these types are less useful, as they would result in more false-positive cases that are likely not relevant to the requirements. Therefore, these 3 types were excluded from the final NLP solution. Compared with the snippet level, this *schema mismatch* caused errors. For instance, *colon rectal cancer* was annotated as *GASTROINTESTINAL* to capture both, but in the NLP implementation, only one RECTAL cancer was counted instead of two cancers to simplify the implementation. This mismatch did not affect the patient-level results but was counted as a snippet-level error.

Table 5. Type of snippet-level errors and counts.

Type of errors	False positive, n	False negative, n	Examples
Annotation error	10	13	A missed annotation
Data input typo[a]	1	5	bladdler ca[b]
Out of vocabulary[a]	2	6	Precancer
Context error[a]	22	11	Possible, colon cancer, died when pt was *5 years old* {{FATHER}}
Ambiguous input	2	3	{{CANCER, COLON}} ileum {{FATHER}}
Schema mismatch[c]	6	10	See above
Total	43	48	N/A[d]

[a]These 3 types of errors are natural language processing (NLP)–caused errors or can be fixed by improving the NLP.

[b]ca: cancer.

[c]This type of error does not need to be fixed.

[d]N/A: not applicable.

NLP-Augmented Algorithm Evaluation Results

The first part of this evaluation compared the NLP-augmented algorithm (using the inclusion configuration) with the structured data algorithm over a stratified sample of 100 patients (50 breast cancer and 50 colorectal cancer, with a 1:2 ratio of cases with agreement versus disagreement between unstructured and structured data). The NLP-augmented algorithm performed better than the structured data algorithm both in precision (0.99, 95% CI 0.96-1.00 vs 0.81, 95% CI 0.65-0.95), recall (0.95, 95% CI 0.90-0.99 vs 0.29, 95% CI 0.19-0.40), and F1 scores (0.97, 95% CI 0.94-0.99 vs 0.43, 95% CI 0.31-0.54).

In the second part of this evaluation, using the whole data set, compared with the original structured FHH entries, NLP augmentation yielded 21,703 (33.6%) additional computable FHH entries, with 8692 (27.9%) entries added owing to the extraction of conditions, 2689 (69.3%) owing to age of onset, and 10,322 (34.9%) owing to family members. With these additional entries extracted by NLP, 1578 (51%) patients met the NCCN criteria for breast cancer genetic testing, 373 (94%) patients met the criteria for colorectal cancer genetic testing, and 1841 (53.8%) additional unique patients met either or both criteria.

Discussion

Principal Findings

This study developed and evaluated an NLP-augmented algorithm to identify patients who met evidence-based criteria for genetic testing of hereditary colorectal and breast cancer. Overall, the proposed automated algorithm offers a promising approach to identifying these patients as an alternative to current clinical workflows, which rely on extensive manual review of patient records. We also demonstrated that compared with structured data alone, an NLP algorithm that focused on the interplay between structured data and associated free-text comments significantly increased the computability of FHH entries and algorithm accuracy. Compared with structured data alone, NLP augmentation led to a 53.8% increase in the number of patients available to compute against the NCCN criteria for genetic testing.

Chen et al recognized the significance of data recorded in the FHH section of an EHR [12]. They characterized the use and contents of the FHH comments field and found that it was used to augment or modify the attributes of the statement (eg, uncertainty and negation) for all 3 types of entities: *family*

member, *condition*, and *age of onset*. However, they did not develop a complete solution for extracting these relationships. In a previous study, we used NLP to extract the disease age of onset from the comments field [10]. In this study, we extended the NLP solution to extract all 3 types of entities and the relations between them. In addition, the algorithm reconciles information from structured and unstructured data to identify patients who meet the NCCN criteria for genetic testing of 2 common hereditary cancers. The study results demonstrated that the NLP-augmented algorithm accurately extracted relevant FHH at the snippet level that combined the structured and comments fields. At the patient level, the algorithm significantly improved the recall and precision of identifying patients who met the NCCN criteria for genetic testing of hereditary breast colorectal cancer.

Compared with previously published studies on FHH extraction using NLP, this study differs significantly in the input data source, types of technical challenges, and ultimate goals. Previous studies have focused primarily on extracting FHH from clinical notes, whereas our approach targets the FHH section of the EHR by combining structured and unstructured data. Complete sentences are typical in the FHH narrative of clinical notes, while single words, phrases, and short sentences are more typical in the FHH comment fields. Consequently, the technical challenges are different. Challenges in extracting FHH from clinical notes include FHH section detection, entity recognition, and relation detection [9,14,15]. In contrast, targeted extraction from the FHH section of the EHR requires reconciliation between structured and unstructured data, as they can be complementary, redundant, or conflicting [12]. In addition, extraction from clinical notes focuses on general FHH extraction, whereas our approach aims to identify patients with a specific clinical purpose. Thus, the NLP performance reported in Table 4 is not directly comparable with that reported in previous studies.

As noted above, the NLP-augmented algorithm can be configured to include or exclude FHH entries with uncertain statements in the free-text comments. The choice of configuration depends on the requirements of specific use cases and available institutional resources. For instance, in a study that aimed to reach out to eligible patients offering genetic testing, a higher priority may have been given to patients who met testing criteria with a higher degree of certainty (ie, excluding uncertain statements) to minimize manual screening efforts. In contrast, if genetic testing outreach is rolled out as usual care, an institution may want to maximize the benefits of genetic testing to as many patients as possible by including uncertain statements. The difference in algorithm performance between the 2 configurations (ie, including vs excluding uncertainty statements) was not significant. Thus, we did not report the results using the exclusion configuration.

The results showed that the NLP-augmented algorithm had significantly higher precision and recall than structured data alone in identifying patients who met the NCCN criteria for genetic testing. This increase was achieved because the comments field provided additional information that can be used to compute the NCCN criteria, including the cancer type (eg, *pancreatic cancer)*, the age of onset (eg, *diagnosed colon cancer, at age 40*), and the affected family member (eg, *paternal aunt*). In addition, information in the comments field can correct inaccurate data in structured fields.

Limitations

This study had several limitations. First, we used data from one EHR at an academic medical center. Therefore, we cannot conclude that the algorithm and study findings are generalizable to other EHRs and health care systems. However, the EHR used in this study is one of the most widely used EHRs in the United States, and other EHR products use similar FHH sections to collect FHH data [12], suggesting that the proposed approach may be adapted to those settings. Second, error analysis demonstrated that certain FHH entries could not be disambiguated based on the available data provided in the FHH section. Future studies could investigate approaches to disambiguate these FHH entries, such as applying NLP to clinical notes or asking patients to confirm through the patient portal.

As the patient-level data set down-sampled the cases in which the 2 algorithms agreed, the difference between the NLP-augmented algorithm and the structured data algorithm was amplified correspondingly. Thus, we did not analyze the statistical differences between the algorithms on this data set. Despite this, the results showed that when these 2 algorithms disagreed with each other, the NLP-augmented algorithm likely received correct answers. In addition, because of the down-sampling, more challenging cases were likely included in the reference data set compared with the original data set. Thus, the actual performance of both algorithms is potentially higher than the scores reported in the section of *NLP-Augmented Algorithm Evaluation Results*.

Although the NLP-augmented algorithm still missed eligible patients, it achieved higher recall than the structured algorithm. Future studies could investigate combining FHH extraction from both FHH sections and clinical notes to further reduce false-negative errors. In addition, other solutions beyond NLP are needed to improve the accuracy and comprehensiveness of the FHH collection in the EHR.

Finally, we investigated only a rule-based solution for the NLP task. Given that the performance was satisfactory and the rule-based approach could be customized quickly for error fixing and future enhancements, we decided that it was not worthwhile to investigate more complex machine learning–based solutions.

Conclusions

This study demonstrated that our NLP solution can accurately extract FHH from both the structured and unstructured fields of the FHH section. Applying this NLP solution to augment the structured data algorithm could improve the precision and recall of identifying patients who meet the NCCN criteria for genetic testing of hereditary breast and colorectal cancer.

Acknowledgments

This research was supported by grants U24CA204800 and U01CA232826 from the National Cancer Institute of the United States. National Institutes of Health and T15LM007124 of the National Library of Medicine.

Conflicts of Interest

None declared.

References

1. Onega T, Beaber EF, Sprague BL, Barlow WE, Haas JS, Tosteson AN, et al. Breast cancer screening in an era of personalized regimens: a conceptual model and National Cancer Institute initiative for risk-based and preference-based approaches at a population level. Cancer 2014 Oct 01;120(19):2955-2964 [FREE Full text] [doi: 10.1002/cncr.28771] [Medline: 24830599]
2. Kahi CJ, Imperiale TF, Juliar BE, Rex DK. Effect of screening colonoscopy on colorectal cancer incidence and mortality. Clin Gastroenterol Hepatol 2009 Jul;7(7):770-5; quiz 711. [doi: 10.1016/j.cgh.2008.12.030] [Medline: 19268269]
3. Armstrong AC, Evans GD. Management of women at high risk of breast cancer. BMJ 2014 Apr 28;348(apr28 26):g2756. [doi: 10.1136/bmj.g2756] [Medline: 24778341]
4. Walter LC, Schonberg MA. Screening mammography in older women: a review. JAMA 2014 Apr 02;311(13):1336-1347 [FREE Full text] [doi: 10.1001/jama.2014.2834] [Medline: 24691609]
5. Nelson HD, Pappas M, Zakher B, Mitchell JP, Okinaka-Hu L, Fu R. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: a systematic review to update the U.S. Preventive services task force recommendation. Ann Intern Med 2014 Feb 18;160(4):255-266. [doi: 10.7326/m13-1684]
6. Daly MB, Pilarski R, Yurgelun MB, Berry MP, Buys SS, Dickson P, et al. NCCN guidelines insights: genetic/familial high-risk assessment: breast, ovarian, and pancreatic, version 1.2020. J Natl Compr Canc Netw 2020 Apr;18(4):380-391 [Referenced with permission from the National Comprehensive Cancer Network, Inc. 2020]. [doi: 10.6004/jnccn.2020.0017] [Medline: 32259785]
7. Gupta S, Provenzale D, Llor X, Halverson AL, Grady W, Chung DC, et al. NCCN guidelines insights: genetic/familial high-risk assessment: colorectal, version 2.2019. J Natl Compr Canc Netw 2019 Sep 01;17(9):1032-1041 [Referenced with permission from the National Comprehensive Cancer Network, Inc. 2020]. [doi: 10.6004/jnccn.2019.0044] [Medline: 31487681]
8. Del Fiol G, Kohlmann W, Bradshaw RL, Weir CR, Flynn M, Hess R, et al. Standards-based clinical decision support platform to manage patients who meet guideline-based criteria for genetic evaluation of familial cancer. JCO Clin Cancer Informatics 2020 Nov(4):1-9. [doi: 10.1200/cci.19.00120]
9. Bill R, Pakhomov S, Chen ES, Winden TJ, Carter EW, Melton GB. Automated extraction of family history information from clinical notes. AMIA Annu Symp Proc 2014;2014:1709-1717 [FREE Full text] [Medline: 25954443]
10. Mowery DL, Kawamoto K, Bradshaw R, Kohlmann W, Schiffman JD, Weir C, et al. Determining onset for familial breast and colorectal cancer from family history comments in the electronic health record. AMIA Jt Summits Transl Sci Proc 2019;2019:173-181 [FREE Full text] [Medline: 31258969]
11. Mehrabi S, Wang Y, Ihrke D, Liu H. Exploring gaps of family history documentation in EHR for precision medicine -a case study of familial hypercholesterolemia ascertainment. AMIA Jt Summits Transl Sci Proc 2016;2016:160-166 [FREE Full text] [Medline: 27570664]
12. Chen E, Melton G, Burdick T, Rosenau P, Sarkar I. Characterizing the use and contents of free-text family history comments in the Electronic Health Record. AMIA Annu Symp Proc 2012;2012:85-92 [FREE Full text] [Medline: 23304276]
13. Taber P, Ghani P, Schiffman J, Kohlmann W, Hess R, Chidambaram V, et al. Physicians' strategies for using family history data: having the data is not the same as using the data. JAMIA Open 2020 Oct;3(3):378-385 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa035] [Medline: 34632321]
14. Shen F, Liu S, Fu S, Wang Y, Henry S, Uzuner O, et al. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 national NLP clinical challenges (n2c2)/open health natural language processing (OHNLP) competition. JMIR Med Inform 2021 Jan 27;9(1):e24008 [FREE Full text] [doi: 10.2196/24008] [Medline: 33502329]
15. Yang X, Zhang H, He X, Bian J, Wu Y. Extracting family history of patients from clinical narratives: exploring an end-to-end solution with deep learning models. JMIR Med Inform 2020 Dec 15;8(12):e22982 [FREE Full text] [doi: 10.2196/22982] [Medline: 33320104]
16. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012 Presented at: Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics; Apr 23 - 27, 2012; Avignon, France URL: https://aclanthology.org/E12-2021
17. Cancer family history annotation schema. GitHub. URL: https://github.com/jianlins/fmx_schema [accessed 2022-03-07]
18. Shi J, Mowery D, Zhang M, Sanders J, Chapman W, Gawron L. Extracting intrauterine device usage from clinical texts using natural language processing. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics

(ICHI). 2017 Presented at: 2017 IEEE International Conference on Healthcare Informatics (ICHI); Aug 23-26, 2017; Park City, UT, USA. [doi: 10.1109/ichi.2017.21]
19. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc 2008 Nov 06;2008:247-251 [FREE Full text] [Medline: 18999129]
20. Shi J, Hurdle JF. Trie-based rule processing for clinical NLP: a use-case study of n-trie, making the ConText algorithm more efficient and scalable. J Biomed Inform 2018 Sep;85:106-113 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.002] [Medline: 30092358]
21. EasyCIE_Hub. GitHub. URL: https://github.com/jianlins/EasyCIE_Hub [accessed 2022-06-14]
22. Austin PC, Tu JV. Bootstrap methods for developing predictive models. Am Statistician 2004 May;58(2):131-137. [doi: 10.1198/0003130043277]
23. Kawamoto K. OpenCDS: an open-source, standards-based, service-oriented framework for scalable CDS. In: Proceedings of the SOA in Healthcare 2011 Conference. 2011 Presented at: SOA in Healthcare 2011 Conference; Jul 13-15, 2011; Hyatt Dulles, Herndon. [doi: 10.1016/j.csi.2020.103468]
24. Kaphingst KA, Kohlmann W, Chambers RL, Goodman MS, Bradshaw R, Chan PA, BRIDGE research team. Comparing models of delivery for cancer genetics services among patients receiving primary care who meet criteria for genetic evaluation in two healthcare systems: BRIDGE randomized controlled trial. BMC Health Serv Res 2021 Jun 02;21(1):542 [FREE Full text] [doi: 10.1186/s12913-021-06489-y] [Medline: 34078380]

Abbreviations

CDS: clinical decision support
EasyCIE: easy clinical information extractor
EHR: electronic health record
FHH: family health history
NCCN: National Comprehensive Cancer Network
NLP: natural language processing
SQL: Structured Query Language

Edited by T Hao; submitted 09.03.22; peer-reviewed by S Fu, J Xia, P Han; comments to author 04.05.22; revised version received 29.06.22; accepted 06.07.22; published 11.08.22.

Please cite as:
Shi J, Morgan KL, Bradshaw RL, Jung SH, Kohlmann W, Kaphingst KA, Kawamoto K, Fiol GD
Identifying Patients Who Meet Criteria for Genetic Testing of Hereditary Cancers Based on Structured and Unstructured Family Health History Data in the Electronic Health Record: Natural Language Processing Approach
JMIR Med Inform 2022;10(8):e37842
URL: https://medinform.jmir.org/2022/8/e37842
doi:10.2196/37842
PMID:35969459

Original Paper

Exploiting Intersentence Information for Better Question-Driven Abstractive Summarization: Algorithm Development and Validation

Xin Wang[1], MSc; Jian Wang[1], PhD; Bo Xu[1], PhD; Hongfei Lin[1], PhD; Bo Zhang[1], PhD; Zhihao Yang[1], PhD

School of Computer Science and Technology, Dalian University of Technology, Dalian, China

Corresponding Author:
Jian Wang, PhD
School of Computer Science and Technology
Dalian University of Technology
No 2 Linggong Road
Dalian, 116023
China
Phone: 86 13604119266
Email: wangjian@dlut.edu.cn

Abstract

Background: Question-driven summarization has become a practical and accurate approach to summarizing the source document. The generated summary should be concise and consistent with the concerned question, and thus, it could be regarded as the answer to the nonfactoid question. Existing methods do not fully exploit question information over documents and dependencies across sentences. Besides, most existing summarization evaluation tools like recall-oriented understudy for gisting evaluation (ROUGE) calculate N-gram overlaps between the generated summary and the reference summary while neglecting the factual consistency problem.

Objective: This paper proposes a novel question-driven abstractive summarization model based on transformer, including a two-step attention mechanism and an overall integration mechanism, which can generate concise and consistent summaries for nonfactoid question answering.

Methods: Specifically, the two-step attention mechanism is proposed to exploit the mutual information both of question to context and sentence over other sentences. We further introduced an overall integration mechanism and a novel pointer network for information integration. We conducted a question-answering task to evaluate the factual consistency between the generated summary and the reference summary.

Results: The experimental results of question-driven summarization on the PubMedQA data set showed that our model achieved ROUGE-1, ROUGE-2, and ROUGE-L measures of 36.01, 15.59, and 30.22, respectively, which is superior to the state-of-the-art methods with a gain of 0.79 (absolute) in the ROUGE-2 score. The question-answering task demonstrates that the generated summaries of our model have better factual constancy. Our method achieved 94.2% accuracy and a 77.57% F1 score.

Conclusions: Our proposed question-driven summarization model effectively exploits the mutual information among the question, document, and summary to generate concise and consistent summaries.

(JMIR Med Inform 2022;10(8):e38052) doi:10.2196/38052

KEYWORDS

question-driven abstractive summarization; transformer; multi-head attention; pointer network; question answering; factual consistency; algorithm; validation; natural language processing

Introduction

Automatic text summarization of natural language aims to summarize the source document to generate a concise and informative description for helping people efficiently and quickly capture the main idea [1,2]. In the biomedical domain, question-driven answer summarization can be particularly useful for people whether they have a biomedical background or not because the generated summary only covers the key information with respect to a specific question and filters out the explanation part [3]. It is different from a factoid question-answering (QA) [4] system. The answer of factoid QA is a phrase or a sentence according to the question, but users prefer the detailed answer including more information to the accurate answer. Summaries for nonfactoid questions [5] should be semantically consistent

and identical with the context. PubMedQA [6] is a novel biomedical nonfactoid QA data set collected from PubMed articles in which the title is a question and can be answered by yes or no. Some related studies [7,8] treat this QA data set as a summarization task and take the conclusion part of the abstract as the answer summary.

Early works put emphasis on query-based summarization approaches [9-11] in which the aim is to extract the sentences relevant to the given query. However, these methods are typically based on semantic relevance from query to context and neglect mutual information at the sentence level, which is helpful for the reasoning or inference process in question-driven summarization. These traditional extractive summarization methods are mainly based on information retrieval methods to select sentences that heavily rely on feature engineering, and the results performance is restricted by pipelines [5,12,13]. Though extractive summarization is more grammatical and coherent, the extractive sentences fail to have a logical connection. In contrast to extractive methods, abstractive methods produce summaries at the word level based on semantic comprehension [8]. Consequently, question-driven abstractive answer summarization is studied to generate the concise and salient short answer, which is also informative for answering the question.

To tackle question-driven abstractive summarization, the answer summary should be highly related to the concerned question. Existing studies [7,8,14] often concentrate on processing the mutual information between the question and document. However, though some sentences are not strongly related to the question, they further explain the central entity in question and affect the expression of the context. Mutual information among answer sentences is underused. Furthermore, it is hard for the recurrent neural network (RNN)–based model to capture the information of long sentences. Existing studies model the sentences separately, which hinders the interaction among sentences. To this end, we propose a novel transformer-based model [15] named Trans-Att that incorporates a two-step attention mechanism to enhance the mutual information both of question to context and sentence over other sentences. A novel multi-view pointer-generator network is proposed to create a condensed and concise summary to better use the question and context information.

Furthermore, a common problem in the practical application of abstractive summarization models is the factual inconsistency [16]. This refers to the phenomenon that the model produces a summary that sometimes distorts and fabricates the facts. Recent studies point out that up to 30% of the generated summaries contain such factual inconsistencies [16,17]. One main reason is that most existing summarization evaluation tools calculate N-gram overlaps between the generated summary and the reference [16]. Though some models make higher scores in token-level metrics like recall-oriented understudy for gisting evaluation (ROUGE) [18], the generated summaries still lack factual correctness. Thus, human evaluation is still the primary method for evaluating the factual consistency. In question-driven answer summarization, generated summaries should be consistent with the context semantically. Wang et al [19] and Durmus et al [20] propose the QA-based factual consistency evaluation metrics QAGS and FEQA separately. They first generate a set of questions about the summary and then use a QA model to answer these questions for evaluation. Because of the characteristics of the PubMedQA data set, the questions are general questions, and they can be answered by yes or no. We use the summaries as the context for the QA task to evaluate the factual consistency.

In this paper, a novel question-driven abstractive summarization based on transformer is proposed, namely Trans-Att, that incorporates a two-step attention mechanism and an overall integration mechanism to summarize the document with respect to the nonfactoid questions. Concretely, the two-step attention mechanism can learn richer structural dependencies among sentences and the relevance of the question and the document. The overall integration mechanism integrates the question, the document, and the correlative summary to generate a summary representation, which allows the model to use the comprehensive information. A novel multi-view pointer network is then proposed by integrating transformer and pointer-generator networks [21] to facilitate copy words from the question or the document to better use the question and context information. Finally, besides question-driven abstractive summarization evaluated by ROUGE, we also assess the model performance by QA task to evaluate the generated summary and whether they are factually consistent with the source document with regard to the question. The effectiveness of this model is empirically validated on the text summarization task and QA task, and achieves state-of-the-art performance on the PubMedQA data set.

The following are our main contributions. First, the novel architecture Trans-Att uses a two-step attention mechanism for better integrating the information in both question to context and sentence over other sentences.

Second, we propose a novel multi-view pointer network to generate tokens through overall integration, which integrates the attentive question, the attentive document, and the correlative summary to generate a summary representation.

Finally, besides ROUGE for automatically evaluating the summarized answers, we conduct a QA task to evaluate the factual consistency.

Methods

Question-Driven Abstractive Summarization

Automatic text summarization is a challenging task in the natural language processing field. It aims to generate simple and coherent essays that comprehensively and accurately reflect the central content of an original document. It can be categorized into two approaches: extractive and abstractive methods. The former method selects a few relevant sentences from the original text, while the latter needs to rephrase and generate a new sentence in which some words are not necessarily present in the original text. In this paper, we focus on abstractive summarization for its potential of summarizing the text more coherently and logically.

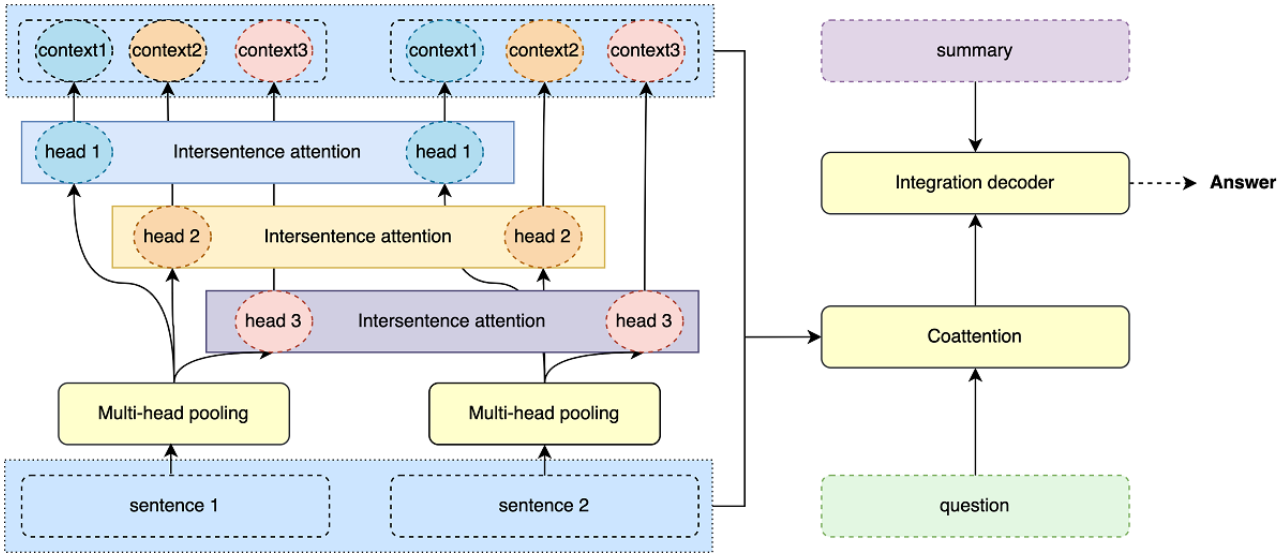
Question-driven summarization is intended to summarize the original document in terms of a specific question, which is different from query-based summarization. In query-based summarization, the query is often a word or a phrase referring to a particular entity [11]. Whereas a question may contain several entities and a specific semantic meaning, and this requires the model to have the reasoning or inference ability to identify the corresponding semantic contents in question-driven summarization [8]. Early query-based summarization methods heavily rely on feature engineering including query-dependent features and query-independent features. The former includes named entity matching and semantic sentence matching, and the latter includes term frequency–inverse document frequency and stop word penalty [1,22]. Recently, some abstractive sequence-to-sequence neural networks have recently been proposed to generate summaries in regard to the given query [10,11]. Some recent works have developed a new method for question-driven summarization [7,8,14] in nonfactoid QA that requires much reasoning and an inference process. However, these methods only model the relation between the question and each sentence, and neglect the mutual information among sentences.

Problem Formulation

For the text summarization task, formally, assume that we have a question $q = \{q_1, ..., q_m\}$ with m words and a source document  containing l^s sentences that have n^s words at most. The task is to generate an answer summary $y = \{y_1, ..., y_n\}$ containing n words. The training goal is to maximize the probability $p(y|q, d)$. The overall architecture of our transformer-based question-driven abstractive answer summarization model is depicted in Figure 1, which consists of three main components: (1) two-step attention mechanism, (2) overall integration mechanism, and (3) multi-view pointer network for generation.

For the QA task, given a question q and an answer summary y, the model should generate an answer $a = \{0,1\}$ indicting yes or no to this question conditioned on the document. We adopted BioBERT [23] as our model to evaluate the factual consistency, which is initialized with bidirectional encoder representation from transformers (BERT) [24] and further pretrained on large-scale biomedical corpora.

Figure 1. Overview of our model.



Encoder

Question Encoder

Let  denote the token embedding indicating the meaning of each token q_i. A special positional encoding pe_i indicates the position of each token within the question sequence. The input of the question encoder I^q is a sequence of embeddings.

A transformer layer is used to encode the question. It reads the question $q = \{q_1, ..., q_m\}$ and computes a hidden representation  , where N_m denotes the length of the question and d is the dimension of the vector. To get a fixed length question representation, H^q is then converted to a vector  by adding all token representations and normalizing it by question length.

Sentence Encoder

Each document is composed of several sentences. Given a document context  , the input of the sentence encoder is the sentences fed one by one. We used sentence position embedding to indicate the order of sentences.

where  is the word embedding of $w_{i,j}$, which is the same word embedding as  ; the position embedding of the token is represented as  , and  denotes the sentence position embedding of  .

I^s then fed into a transformer encoder to represent the sentence as a sequence of hidden vectors by:

The hidden representation of a document is represented as  and a sentence vector  , where $N^s = l^s \times n^s$.

Two-step Attention Mechanism

Intersentence Attention

Inspired by Liu and Lapata [25], we used an intersentence attention mechanism to model the dependencies across multiple sentences, where each sentence can attend to other sentences. We used a weighted-pooling operation to obtain a fixed-length sentence representation so that the diversity of each sentence representation is increased. Through a *multi-head pooling mechanism* [25], each token can attend to other tokens by calculating weight distributions. Sentences can be encoded flexibly in different subspaces.

The output representation  of the last transformer encoder layer for token $w_{i,j}$ is denoted as $x_{i,j}$ as the input. For each sentence  and for head $z \in \{1, ..., n_{head}\}$, we first conducted a linear transformation to obtain the attention scores  and value vectors  . The probability distribution  was then calculated within the sentence.

where  and  are weights. $d_{head} = d / n_{head}$ is the dimension of each head.

Based on the probability distributions and value vectors, we conducted a weighted summation followed by another linear formation and layer normalization. Different vector  encodes sentences in a different subspace.

where  is the weight. Because of the flexibility of combining multiple heads, each sentence has multiple attention distribution and focuses on different views of input.

Dependencies among multiple sentences can be modeled by the intersentence attention that is similar to self-attention. Intersentence attention computes the distribution of attention so that each sentence attends to other sentences.

where  are query, key, and value vectors, respectively. Through a self-attention calculation,  is obtained to represent the sentence vector that gathers the information of other sentences. l^s is the number of input sentences.

We then concatenate all context vectors and pass through a linear layer with weight  to update token representations by adding c_i to each token vector $x_{i,j}$. We then pass it through a two-layer multilayer perception, taking *gelu* as the activation function [26]. Next, we pass the summation of $x_{i,j}$ and $g_{i,j}$ to a layer normalization. In this way, each sentence collects information from other sentences represented as  .

Coattention

Coattention is the second attention mechanism aimed at exploiting the pairwise mutual information between the question and the context.

We further used an additive attention [27] to obtain the distribution of document sentences that highly coincides with the question and then combines the question and question-related sentences to get their comprehensive representation  by:

where *MLP* is the same as mentioned before.  are trainable parameters.

Integration Decoder

When given the first $t - 1$ tokens in the summary $y_1, ..., y_n$, the integration decoder incorporates the question and the document into the summary through an overall integration mechanism. The purpose is to predict the representation of the $t - th$ token and transmit it to the pointer network.

Overall Integration

Inspired by gated recurrent units [28], we designed an *integration* gate (z) to integrate the question-document and summary, which enables summary tokens at different times to merge information in different levels. Multi-head attention is then used to capture the information in the fused representation,  , and obtain s^y, which is a correlative summary.  is the vector representation of the input summary.

To reinforce the understanding of the question and document of the decoder, s^y is used to compute attention with the question and the document, and obtain representations s^q and s^s.

$$s^q = Multi - headAttention\ (s^y, H^q, H^q) \quad \textbf{(23)}$$

$$s^s = Multi - headAttention\ (s^y, H^{s\prime}, H^{s\prime}) \quad \textbf{(24)}$$

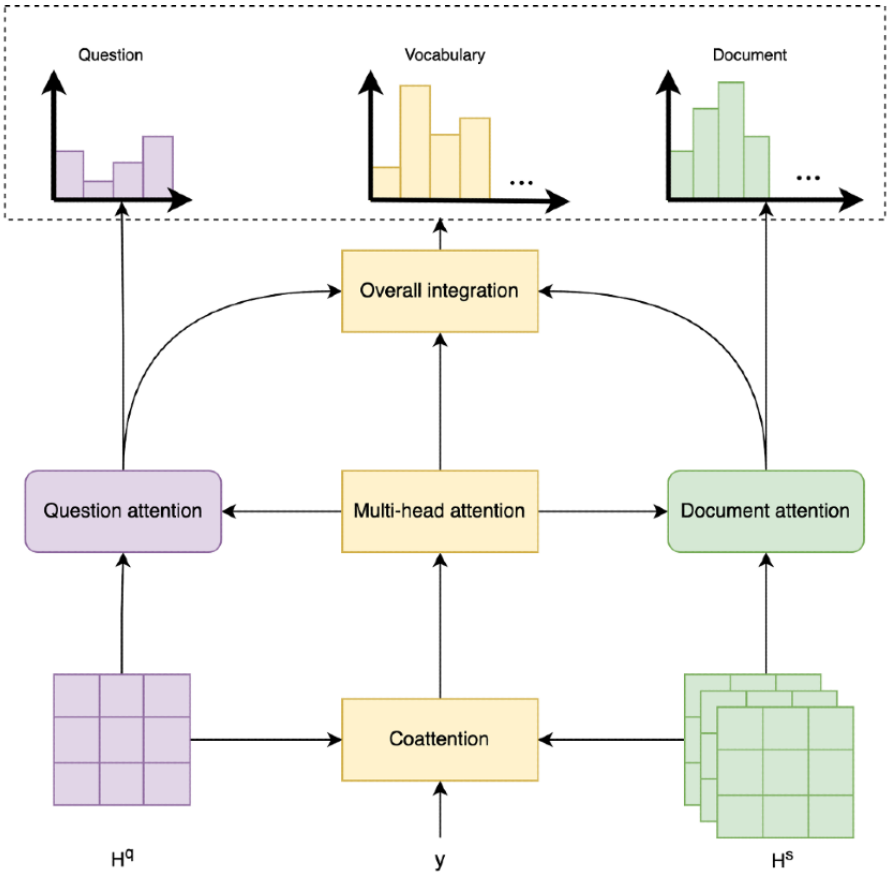
Next, similar to equation 20, the predicted representation o^y is obtained to integrate the attentive question, the attention document, and the correlative summary by using the *integration* gate.

where * is denoted as ^{q} or ^{s}.

Multi-View Pointer Network

To improve the probability of generating corresponding tokens from the question and the document, a novel multi-view pointer network is proposed based on multi-head attention as shown in Figure 2.

Figure 2. Multi-view pointer network. H^q: hidden representation of question; y: hidden representation of the input summary; H^s: hidden representation of document.



Question Tokens

We computed the attention weights β^q through multiple attention weights in the multi-head attention.

Where f_β means a function of getting multiple attentions in the multi-head attention.  is the weight, where n^{head} is the number of heads. β^q can be treated as the probability distribution over the question words. It can be represented as  .

Document Tokens

The distribution of the document that is relevant to the question can be served as a global distribution over each decoding step. β^s can be calculated similar to equation 27, which can be considered a local distribution at each decoding step. Thus, the distribution over the document can be calculated by:

Vocabulary Tokens

The predicted representation from the overall integration decoder is used to calculate the probability distribution p^v over the fixed vocabulary through a *softmax* layer; W^v is the weight from the word embeddings.

The final probability distribution y_t to predict can be formulated from three aspects of word distributions as:

$$P(y_t \mid q, d, y < t) = softmax\,(W_\gamma o^y + b_\gamma) \cdot [p^v, p^q, p^s] \quad \textbf{(31)}$$

Loss Function

The main training objective is to minimize the negative log likelihood between the reference summary and the predicted summary. Thus, Trans-Att can be trained by minimizing the objective.

Question-Answering Model

BERT [24] has already been used in QA tasks. We fine-tuned BioBERT [23] as a baseline. We fed PubMedQA questions and

corresponding texts that could be contexts, reference long answers, contexts and long answers, or generated summaries for comparison, separated by special [SEP] token, to the model. We take the special embedding [cls] from the last layer and use a *softmax* function to predict the final label that could be yes or no. The general loss was trained by minimizing the cross-entropy between the predicted labels and the true label distribution.

Results

Data Set

We evaluated our model on the nonfactoid QA data set PubMedQA [6]. PubMedQA is a novel biomedical data set aiming at answering academic questions and has substantial instances with some expert annotations. Each instance is composed of a question that is a general question, a context that is the structured abstract without its conclusion, a long answer that is the conclusion of the abstract in terms of the question, and a final answer yes/no for the general question that summarizes the conclusion and can be used for the QA task. The statistics of the PubMedQA data set are shown in Table 1.

We adopted ROUGE-1, ROUGE-2, and ROUGE-L to automatically evaluate the summarized answers in the question-driven abstractive summarization task. The main metrics of the QA task are accuracy and macro-F1 under a reasoning-free setting in which the generated summary is added in the input.

Table 1. Statistics of the PubMedQA data set.

Task data set	Training, n	Development, n	Test, n
QA[a] pairs	169,000	21,000	21,000
Average question length (word count)	16.3	16.4	16.3
Average document length (word count)	238	238	239
Average summary length (word count)	41.0	41.0	40.9
Average number of sentences	9.32	9.31	9.33

[a]QA: question-answering.

Experimental Settings

ParlAI [29] was implemented in our model as the code framework. The dimensions of word embedding size and hidden size were both 256. The text was encoded by byte-pair encoding [30], and the embedding matrix was initialized with fastText. Both encoder and decoder layers of transformer-based models were 5, with feed-forward hidden size 512 and attention head 4 for all layers. The optimizer was Adam [31] with an initial learning rate of 0.0005. We also applied the inverse square root learning schedule over the 5k warm-up dates. The dropout rate was set to 0.2, and gradient clipping was used with a maximum gradient norm of 0.1. Label smoothing of the value 0.1 was used for summary generation. We used beam search in the generation process with beam size 2 and adopted 3-gram blocking.

Comparative Methods

We report the performance of our proposed model in comparison with several baselines and state-of-the-art methods based on different methodologies, including extractive summarization, abstractive summarization, query-based summarization, and question-driven abstractive summarization.

Two unsupervised extractive methods were used. LEAD3 is a simple but effective extractive summarization baseline that concatenates the first two sentences and the last sentence without question information. Maximal marginal relevance is an information retrieval model used to calculate the similarity between the text and the researched document for extractive summarization.

Three widely adopted abstractive methods were adopted for comparison. Sequence-to-sequence model with attention [27] is a simple encoder-decode model with attention based on RNN without respect to the question. Pointer-generator network [21] is a hybrid pointer-generator architecture with coverage based on a neural sequence-to-sequence model for abstractive text summarization. Transformer [15] implements the state-of-the-art encoder-decoder framework based on multi-head attention without access to the question.

There were two query-based abstractive summarization methods used for comparison. The soft long short-term memory–based diversity attention model (SD_2) [10] adds a query attention mechanism to a sequence-to-sequence model. It learns to pay attention to different parts of the query at different time steps. Query-based summarization using neural networks (QS) [11] incorporates question information into the pointer-generator network with the use of the vanilla attention mechanism.

Finally, we implemented two of the latest question-driven answer summarization models for comparison. Hierarchical and sequential context modeling [7] is a hierarchical compare-aggregate method used to integrate the interaction between the question and the document into final document representation at both the word level and sentence level. Multi-hop selective generator (MSG) [8] models the relevance between question and sentences by leveraging a humanlike multi-hop reasoning process for question-driven summarization, in which the most related sentences are given higher weights.

Experimental Results

The experimental results of question-driven summarization in terms of ROUGE scores and QA with respect to accuracy and macro-F1 scores are presented in Tables 2 and 3. Both ROUGE scores and metrics of QA show that our model achieved

competitive performance in comparison with state-of-the-art question-driven summarization methods.

Compared with traditional text summarization, there was limited improvement for query-based summarization methods (SD_2 and QS), indicating that the question information was not sufficiently used. There was a noticeable margin, about 0.79 for ROUGE-2, higher than the current state-of-the-art model (MSG). This indicates that the model benefits from the information provided by mutual information between question and document, and among sentences. We noticed that the ROUGE-1 score of our model was lower than MSG. One possible explanation is that the length of the generated summary of MSG was longer than that of our model. Considering the characteristic of ROUGE-1 that measures the word overlap between the reference summary and the predicted summary, the longer summary has more possibility of generating words that appeared before.

As for the QA result, we observed that if using the original answer summary, BioBERT achieves good enough scores. If the input answer summary can correctly answer the question, it is consistent to the original semantics. Thus, evaluating the factual consistency by a QA task is feasible. Suppose that we feed the context without long answer information to the model, which is under the reasoning-required setting; the result is comparatively lower because the reasoning and inference process is crucial in answering the question if the answer is not directly available. We treated the long answer as the summary, and its quality influenced the factual consistency. It was observed that there is still a big gap between the generated summary and the reference summary, which leaves room for improvement.

Overall, the difference upon accuracy measurement was not significant by a narrow margin because of the imbalanced distribution of labels (92.8% yes vs 7.2% no). The F1 score was significant and representative, and our model achieved the best *F* score of 77.57%. The results show that the extractive methods performed better than the abstractive methods. We speculate that extractive summarization approaches directly copy from the source context. However, it is worth noting that the extractive methods have an upper bound, and they barely exceed the performance when given the whole context. There is substantial potential for abstractive approaches. Future work should explore the reasoning ability of abstractive methods.

Table 2. Comparison with related works of question-driven summarization task.

Methods	Types	With question	ROUGE[a]-1 (%)	ROUGE-2 (%)	ROUGE-L (%)
LEAD3	Extractive	No	30.94	9.79	25.89
MMR[b]	Extractive	No	29.69	9.50	24.10
S2SA[c]	Abstractive	No	32.40	11.00	27.30
PGN[d]	Abstractive	No	32.89	11.51	28.10
Transformer	Abstractive	No	32.38	11.34	26.32
SD_2[e]	Abstractive	Query based	32.33	10.52	26.01
QS[f]	Abstractive	Query based	32.60	11.10	26.70
HSCM[g]	Extractive	Question driven	32.34	10.07	25.98
MSG[h]	Abstractive	Question driven	*37.20* [i]	14.80	30.20
Trans-Att (ours)	Abstractive	Question driven	36.01	*15.59*	*30.22*

[a]ROUGE: recall-oriented understudy for gisting evaluation.

[b]MMR: maximal marginal relevance.

[c]S2SA: sequence-to-sequence model with attention.

[d]PGN: pointer-generator network.

[e]SD_2: soft long short-term memory–based diversity attention model.

[f]QS: query-based summarization using neural networks.

[g]HSCM: hierarchical and sequential context modeling.

[h]MSG: multi-hop selective generator.

[i]Italics indicate the best result.

Table 3. Comparison with related work for question-answering task.

Methods	Accuracy (%)	F1 (%)
LEAD3	93.80	67.06
MMR[a]	*94.85* [b]	75.69
S2SA[c]	91.89	63.81
PGN[d]	91.93	64.42
Transformer	94.18	69.59
SD_2[e]	94.34	69.30
HSCM[f]	93.78	76.48
MSG[g]	93.68	73.27
Trans-Att (ours)	94.20	*77.57*
Majority	92.76	48.12
Context	96.50	84.65
Long answer	99.04	96.18
Context + long answer	99.20	96.86

[a]MMR: maximal marginal relevance.

[b]Italics indicate the best result.

[c]S2SA: sequence-to-sequence model with attention.

[d]PGN: pointer-generator network.

[e]SD_2: soft long short-term memory–based diversity attention model.

[f]HSCM: hierarchical and sequential context modeling.

[g]MSG: multi-hop selective generator.

Ablation Study

To examine the contributions of our proposed modules, namely, intersentence attention, coattention, overall integration, and multi-view pointer network, we ran an ablation study. The experimental results are shown in Table 4.

Overall, all the modules contributed to the final performance to some extent. The accuracy score was not significant compared with the F1 score because of the imbalanced distribution of labels. When the coattention was discarded, the performance of the model dropped substantially, which indicates that it plays a more important role in exploiting the pairwise mutual information between the question and the document sentences. Besides, applying intersentence attention also improved the performance, which indicates that it is not enough to only consider the question-related information. Interrelation among sentences is also worth paying attention to. The decrease on F1 was most significant, which demonstrates the effects of the two-step attention mechanism. Overall integration reinforces the understanding of the model upon the question and the document indicated by a noticeable decrease in F1. Because of the biomedical characteristic of PubMedQA, the out-of-vocabulary problem is much more severe. The ablation study results validated the importance of the multi-view pointer network.

Table 4. An ablation study for our model.

Methods	ROUGE[a]-1	ROUGE-2	ROUGE-L	Accuracy (%)	F1 (%)
Trans-Att	36.01	15.59	30.22	94.20	77.57
Intersentence attention	34.65	13.92	28.07	93.87	73.13
Coattention	34.05	13.61	26.50	93.40	70.62
Overall integration	34.28	14.26	28.63	94.53	72.37
Multi-view pointer network	35.16	13.98	29.32	94.39	75.67

[a]ROUGE: recall-oriented understudy for gisting evaluation.

Case Study

In Figure 3, we show the summaries generated by the proposed method and some baseline methods for comparison, and visualize the sources of the summaries with colors. The context underlined and highlighted with green was used by Trans-Att to generate the summary, which contains more information than in the reference summary. By comparison, we observed that

Trans-Att not only successfully exploits the intersentence information with useful information but also uses the question information in understanding semantic content; pointer-generator network generates an irrelevant summary, which proves the importance of the question information; SD_2 fails to capture the core argument, resulting in repeating the question and paying attention to wrong information; the final answer demonstrates the validity in evaluating factual consistency by QA task (although SD_2 gives the right final answer, there is still a semantic mismatch because the first sentence is essentially the same as the question); and the bottom example demonstrates that there are limitations to the yes/no questions, the answer of which depends partly on clues of negative pronouns. Future work will consider increasing the diversity of the QA task.

Figure 3. Case study from PubMedQA (the bottom example omits the context; final answer is in parentheses). MSG: multi-hop selective generator; PGN: pointer-generator network; QS: query-based summarization using neural networks; SD_2: soft long short-term memory–based diversity attention model; HELLP: hemolysis, elevated liver enzymes, and low platelets counts syndrome.

Question: Does functional brain maturation assessed during early life correlate with anatomical brain maturation at term-equivalent age in preterm infants?
Document: Amplitude-integrated electroencephalogram (aEEG) is a reliable monitoring tool for electrocortical activity with good predictive value in preterm infants. Magnetic resonance imaging (MRI) is a good neuroimaging tool to detect brain lesions and to evaluate brain maturation. We hypothesized that early aEEG measures, recorded over the first 3 d of life in very preterm infants, correlate with brain maturation and injury score assessed by conventional MRI at term-equivalent age. Thirty-nine infants born at a mean (range) gestational age (GA) of 29.5 (27.0-31.9) wk and birth weight 1,230 (680-2,020) g had continuous aEEG during the first postnatal 72-84 h. aEEG maturity scores and average maximum and minimum amplitudes were evaluated. Conventional brain MRI was performed at 41.2 (37.1-44.1) wk postmenstrual age (PMA) on a 3T GE system and scored qualitatively for injury and maturation. The average aEEG total maturity score and its cycling subscore were positively and significantly associated with the total MRI maturation score after adjustment for GA, morphine sedation, and PMA at MRI examination. No association was found between the aEEG measures and the MRI injury scores.
Reference Summary: Early aEEG maturity seems to relate to structural MRI brain maturation at term-equivalent age in preterm infants. **(YES)**
Trans-Att: The aEEG measures recorded over the first 3 d of life are associated with anatomical brain maturation at term-equivalent age in preterm infants. **(YES)**
MSG: Our results suggest that average aEEG measures are associated with anatomical brain maturation at term-equivalent age. **(YES)**
SD_2: Functional brain maturation assessed during early life correlates with anatomical brain maturation at term-equivalent age in very preterm infants. These findings support the use of MRI as an adjunct to conventional brain MRI in preterm infants. **(YES)**
PGN: These results suggest that early aEEG maturation is a reliable marker for electrocortical activity in very preterm infants. **(NO)**

Question: Is history of preeclampsia associated with an increased risk of thyroid dysfunction?
Reference Summary: Preeclampsia and/or HELLP syndrome are not associated with an increased risk of thyroid dysfunction in later life. **(NO)**
Trans-Att: History of preeclampsia and/or HELLP syndrome is not associated with an increased risk of thyroid dysfunction. **(NO)**
MSG: The prevalence of primary thyroid dysfunction and antibodies to thyroid peroxidase was not significantly different between the two groups. **(NO)**
QS: History of preeclampsia is associated with an increased risk of thyroid dysfunction. Further studies are needed to clarify the role of thyroid peroxidase in the development of thyroid dysfunction in women with a history of preeclampsia. **(YES)**
PGN: Our results suggest that the history of preeclampsia and/or HELLP syndrome is not associated with an increased risk of thyroid dysfunction. **(NO)**

Novel N-Grams

We also analyzed the output of abstractive models by calculating the proportion of novel n-grams that appear in the summaries but not in the source texts. Table 5 shows that summaries of our model account for a lower rate of novel n-grams than the reference summaries, indicating the quality of abstraction. We observed that the traditional abstractive approach (pointer-generator network), copies more phrases, perhaps because it generates more words from the context without being question driven, which increases the probability of unrelated information being selected. Note that MSG produces novel n-grams more frequently. However, it may contain the factual inconsistency problem in generating new words.

Table 5. Proportion of novel n-grams.

Methods	1 grams (%)	2 grams (%)	3 grams (%)	4 grams (%)
Trans-Att	11.00	47.82	67.12	79.38
MSG[a]	13.43	54.66	74.13	85.01
PGN[b]	16.29	43.73	58.38	69.14
Refrence	27.83	72.11	87.17	93.55

[a]MSG: multi-hop selective generator.

[b]PGN: pointer-generator network.

Discussion

Conclusions

In this paper, a novel transformer-based question-driven abstractive summarization model was proposed to generate concise and consistent summaries for nonfactoid QA. A two-step attention mechanism was proposed to exploit the mutual information both of the question to context and the sentence over other sentences. We used the overall integration mechanism and the novel pointer network to better integrate and use information of the question, document, and summary. We conducted a QA task to evaluate the factual consistency between the generated summary and the reference summary. Experimental results demonstrate that our proposed model achieves comparable performance to the state-of-the-art methods.

Future Work

Due to the insufficiency of the data set quantity, we were limited to conducting experiments on PubMedQA. We are looking forward to conducting more persuasive experiments when the insufficiency is lifted. As for the evaluation of the factual consistency, we can also incorporate human expertise to further enhance the credibility of the proposed QA metric. Hopefully, our method can provide some inspiration in the summarization task.

Acknowledgments

The publication of this paper is funded by grants from the Natural Science Foundation of China (62006034) and Natural Science Foundation of Liaoning Province (2021-BS-067)

Authors' Contributions

XW and BZ completed the experiments and wrote the paper. JW and BX provided theoretical guidance and revision of the paper. HL, ZY, and BX contributed to the algorithm design.

Conflicts of Interest

None declared.

References

1. Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey. Artif Intelligence Rev 2016 Mar 29;47(1):1-66. [doi: 10.1007/s10462-016-9475-9]
2. Huang D, Cui L, Yang S, Bao G, Wang K, Xie J, et al. What have we achieved on text summarization? 2020 Nov Presented at: 2020 Conference on Empirical Methods in Natural Language Processing; November 2020; Online p. 446-469. [doi: 10.18653/v1/2020.emnlp-main.33]
3. Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. Sci Data 2020 Oct 02;7(1):322. [doi: 10.1038/s41597-020-00667-z] [Medline: 33009402]
4. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. 2016 Nov Presented at: 2016 Conference on Empirical Methods in Natural Language Processing; November 2016; Austin, TX p. 2383-2392. [doi: 10.18653/v1/d16-1264]
5. Song H, Ren Z, Liang S, Li P, Ma J, de Rijke M. Summarizing answers in non-factoid community question-answering. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017 Feb 2 Presented at: WSDM '17; February 6-10, 2017; Cambridge, United Kingdom p. 405-414. [doi: 10.1145/3018661.3018704]
6. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: a dataset for biomedical research question answering. 2019 Nov Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; November 2019; Hong Kong, China p. 2567-2577. [doi: 10.18653/v1/d19-1259]
7. Deng Y, Zhang W, Li Y, Yang M, Lam Y, Shen Y. Bridging hierarchical and sequential context modeling for question-driven extractive answer summarization. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and

Development in Information Retrieval. 2020 Jul 25 Presented at: SIGIR '20; July 25-30, 2020; Virtual event, China p. 1693-1696. [doi: 10.1145/3397271.3401208]
8. Deng Y, Zhang W, Lam W. Multi-hop inference for question-driven summarization. 2020 Nov Presented at: 2020 Conference on Empirical Methods in Natural Language Processing; November 2020; Online p. 6734-6744. [doi: 10.18653/v1/2020.emnlp-main.547]
9. Cao Z, Li W, Li S, Wei F, Li Y. AttSum: joint learning of focusing and summarization with neural attention. In: Proceedings of COLING 2016. 2016 Dec Presented at: 26th International Conference on Computational Linguistics: Technical Papers; December 2016; Osaka, Japan p. 546-556 URL: https://aclanthology.org/C16-1053
10. Nema P, Khapra M, Laha A, Ravindran B. Diversity driven attention model for query-based abstractive summarization. 2017 Jul Presented at: 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 2017; Vancouver, Canada p. 1063-1072. [doi: 10.18653/v1/p17-1098]
11. Hasselqvist J, Helmertz N, Kågebäck M. Query-based abstractive summarization using neural networks. arXiv Preprint posted online on December 17, 2017. [FREE Full text]
12. Wang L, Raghavan H, Cardie C, Castelli V. Query-focused opinion summarization for user-generated content. In: Proceedings of COLING 2014. 2014 Aug Presented at: 25th International Conference on Computational Linguistics: Technical Papers; August 2014; Dublin, Ireland p. 1660-1669 URL: https://aclanthology.org/C14-1157
13. Yulianti E, Chen R, Scholer F, Croft WB, Sanderson M. Document summarization for answering non-factoid queries. IEEE Trans Knowledge Data Eng 2018 Jan 1;30(1):15-28. [doi: 10.1109/tkde.2017.2754373]
14. Deng Y, Lam W, Xie Y, Chen D, Li Y, Yang M, Shen. Joint learning of answer selection and answer summary generation in community question answering. 2020 Apr 03 Presented at: The Thirty-Fourth AAAI Conference on Artificial Intelligence; February 7-12, 2020; New York, NY p. 7651-7658. [doi: 10.1609/aaai.v34i05.6266]
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Dec 4 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA p. 6000-6010 URL: https://dl.acm.org/doi/10.5555/3295222.3295349
16. Huang Y, Feng X, Feng X, Qin B. The factual inconsistency problem in abstractive text summarization: a survey. arXiv Preprint posted online on May 10, 2021. [FREE Full text] [doi: 10.48550/arXiv.2104.14839]
17. Kryscinski W, McCann B, Xiong C, Socher R. Evaluating the factual consistency of abstractive text summarization. 2020 Nov Presented at: 2020 Conference on Empirical Methods in Natural Language Processing; November 16-20, 2020; Online p. 9332-9346. [doi: 10.18653/v1/2020.emnlp-main.750]
18. Lin CY. ROUGE: a package for automatic evaluation of summaries. 2004 Jul Presented at: Text Summarization Branches Out; July 25-26, 2004; Barcelona, Spain p. 74-81 URL: https://aclanthology.org/W04-1013/
19. Wang A, Cho K, Lewis M. Asking and answering questions to evaluate the factual consistency of summaries. 2020 Jul Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Online p. 5008-5020. [doi: 10.18653/v1/2020.acl-main.450]
20. Durmus E, He H, Diab M. FEQA: a question answering evaluation framework for faithfulness assessment in abstractive summarization. 2020 Jul Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Online p. 5055-5070. [doi: 10.18653/v1/2020.acl-main.454]
21. See A, Liu PJ, Manning CD. Get to the point: summarization with pointer-generator networks. 2017 Jul Presented at: 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 30-August 4, 2017; Vancouver, Canada p. 1073-1083. [doi: 10.18653/v1/p17-1099]
22. Ouyang Y, Li W, Li S, Lu Q. Applying regression models to query-focused multi-document summarization. Inf Processing Manage 2011 Mar;47(2):227-237. [doi: 10.1016/j.ipm.2010.03.005]
23. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
24. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Jun Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
25. Liu Y, Lapata M. Hierarchical transformers for multi-document summarization. 2019 Jul Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy p. 5070-5081. [doi: 10.18653/v1/p19-1500]
26. Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). arXiv Preprint posted online on July 8, 2020. [FREE Full text]
27. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2015 May Presented at: 3rd International Conference on Learning Representation; May 7-9, 2015; San Diego, CA. [doi: 10.48550/arXiv.1409.0473]
28. Cho K, Van MB, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder--decoder for statistical machine translation. 2014 Oct Presented at: 2014 Conference on Empirical Methods in Natural Language Processing; October 26-28, 2014; Doha, Qatar p. 1724-1734. [doi: 10.3115/v1/d14-1179]

29. Miller A, Feng W, Batra D, Bordes A, Fisch A, Lu J, et al. ParlAI: a dialog research software platform. 2017 Sep Presented at: 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; September 7-11, 2017; Copenhagen, Denmark p. 79-84. [doi: 10.18653/v1/D17-2014]
30. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. 2016 Aug Presented at: 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); August 7-12, 2016; Berlin, Germany p. 1715-1725. [doi: 10.18653/v1/p16-1162]
31. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2015 May Presented at: 3rd International Conference on Learning Representation; May 7-9, 2015; San Diego, CA. [doi: 10.48550/arXiv.1412.6980]

Abbreviations

BERT: bidirectional encoder representation from transformers
MSG: multi-hop selective generator
QA: question answering
QS: query-based summarization using neural networks
RNN: recurrent neural network
ROUGE: recall-oriented understudy for gisting evaluation
SD_2: soft long short-term memory–based diversity attention model

Edited by T Hao; submitted 17.03.22; peer-reviewed by D Zhao, C Sun; comments to author 15.05.22; revised version received 26.05.22; accepted 10.06.22; published 15.08.22.

Please cite as:
Wang X, Wang J, Xu B, Lin H, Zhang B, Yang Z
Exploiting Intersentence Information for Better Question-Driven Abstractive Summarization: Algorithm Development and Validation
JMIR Med Inform 2022;10(8):e38052
URL: https://medinform.jmir.org/2022/8/e38052
doi:10.2196/38052
PMID:35969463

Original Paper

Synergy Between Public and Private Health Care Organizations During COVID-19 on Twitter: Sentiment and Engagement Analysis Using Forecasting Models

Aditya Singhal[1*], MSc; Manmeet Kaur Baxi[1*], MSc; Vijay Mago[1], PhD

Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Aditya Singhal, MSc
Department of Computer Science
Lakehead University
955 Oliver Rd
Thunder Bay, ON, P7B 5E1
Canada
Phone: 1 807 709 9571
Email: asinghal@lakeheadu.ca

Abstract

Background: Social media platforms (SMPs) are frequently used by various pharmaceutical companies, public health agencies, and nongovernment organizations (NGOs) for communicating health concerns, new advancements, and potential outbreaks. Although the benefits of using them as a tool have been extensively discussed, the online activity of various health care organizations on SMPs during COVID-19 in terms of engagement and sentiment forecasting has not been thoroughly investigated.

Objective: The purpose of this research is to analyze the nature of information shared on Twitter, understand the public engagement generated on it, and forecast the sentiment score for various organizations.

Methods: Data were collected from the Twitter handles of 5 pharmaceutical companies, 10 US and Canadian public health agencies, and the World Health Organization (WHO) from January 1, 2017, to December 31, 2021. A total of 181,469 tweets were divided into 2 phases for the analysis, before COVID-19 and during COVID-19, based on the confirmation of the first COVID-19 community transmission case in North America on February 26, 2020. We conducted content analysis to generate health-related topics using natural language processing (NLP)-based topic-modeling techniques, analyzed public engagement on Twitter, and performed sentiment forecasting using 16 univariate moving-average and machine learning (ML) models to understand the correlation between public opinion and tweet contents.

Results: We utilized the topics modeled from the tweets authored by the health care organizations chosen for our analysis using nonnegative matrix factorization (NMF): c_{umass}=–3.6530 and –3.7944 before and during COVID-19, respectively. The topics were chronic diseases, health research, community health care, medical trials, COVID-19, vaccination, nutrition and well-being, and mental health. In terms of user impact, WHO (user impact=4171.24) had the highest impact overall, followed by public health agencies, the Centers for Disease Control and Prevention (CDC; user impact=2895.87), and the National Institutes of Health (NIH; user impact=891.06). Among pharmaceutical companies, Pfizer's user impact was the highest at 97.79. Furthermore, for sentiment forecasting, autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) models performed best on the majority of the subsets of data (divided as per the health care organization and period), with the mean absolute error (MAE) between 0.027 and 0.084, the mean square error (MSE) between 0.001 and 0.011, and the root-mean-square error (RMSE) between 0.031 and 0.105.

Conclusions: Our findings indicate that people engage more on topics such as COVID-19 than medical trials and customer experience. In addition, there are notable differences in the user engagement levels across organizations. Global organizations, such as WHO, show wide variations in engagement levels over time. The sentiment forecasting method discussed presents a way for organizations to structure their future content to ensure maximum user engagement.

(JMIR Med Inform 2022;10(8):e37829) doi:10.2196/37829

KEYWORDS

social media; health care; Twitter; content analysis; user engagement; sentiment forecasting; natural language processing; public health; pharmaceutical; public engagement

Introduction

Background

Social media platforms (SMPs), such as Twitter, Facebook, and Reddit, are commonly used by people to access health information. In the United States, 8 in 10 internet users access health information online, and 74% of these use SMPs. Meanwhile, public health agencies and pharmaceutical companies often use social media to engage with the public [1]. SMPs significantly contribute to the community by providing a communication platform for the public, patients, and health care professionals (HCPs) to talk about health concerns, eventually leading to better outcomes [2]. Additionally, SMPs also function as a medium to motivate patients by promoting health care education and providing the latest information to the community [1]. Analyzing social media content in the health care domain can reveal important dimensions, such as audience reach (eg, followers and subscribers), post source (eg, pharmaceutical companies, public health agencies), and post interactivity (eg, number of likes, retweets) [3]. A recent study discussed a machine learning (ML) approach to examining COVID-19 on Twitter [4]. Although it identifies discussion themes, there is no research on understanding the content shared by public health agencies and private organizations.

Related Works

The positive impacts of using SMPs by patients and HCPs have been previously discussed [5]. Patients feel empowered and develop positive relationships with their HCPs. For instance, Ventola [1] discussed SMPs as a tool to share and promote healthy habits, share information, and interact with the public. Li et al [6] presented an analysis of social media's impact on the public. Their research discusses public perceptions of health-related content being classified as true, debatable, or false; the study shows that people have a strong tendency to adopt collective opinions while sharing health-related statements on social media.

There are different topic-clustering and content analysis techniques available to identify the characteristics of stakeholders (eg, pharmaceutical companies' tweets for drug information) on SMPs [7,8]. A previous study presented an overview of techniques used for sentiment analysis in health care [9]. The researchers discuss multiple lexicon-based and ML-based approaches. The previous discussion on pharmaceutical companies has focused on COVID-19 vaccine–related public opinions [10,11]. Using latent dirichlet allocation (LDA) and valence aware dictionary and sentiment reasoner (VADER), researchers have examined topics, trends, and sentiments over time [10].

Prior research work has also focused on the response of G7 leaders during COVID-19 on Twitter [12,13]. The research classified viral tweets into appropriate categories, the most common being *informative*. Furthermore, researchers have recently presented a discussion on the harms and benefits of using Twitter during COVID-19 [14]. An epidemiological study conducted in 2020 investigated the news-sharing behavior on Twitter. Although it concluded that tweets that include news articles sharing pandemic information are popular, they cannot substitute public health agencies, organizations, or HCPs [15]. In addition, the study of public sentiments via artificial intelligence (AI) can provide a way to frame public health policies [16].

COVID-19 led to a rapid change in public sentiments over a short span of time [17]. People expressed sentiments of joy and gratitude toward good health and sadness and anger at the loss of life and stay-at-home orders [17,18]. Understanding public perceptions toward health-related content is important. Although the majority of people have a positive attitude toward social media, some feel more attention is required to promote the credibility of shared information [19]. Attempts have been made to capture peoples' reactions to the pandemic; however, they are limited in scope. One study investigated the concerns originating toward public health interventions in North America via topic modeling [20], while another examined the role of beliefs and susceptibility information in public engagement on Twitter [21]. Statistical analysis also shows that health care organizations have to come forward to engage more with consumers [22]. The importance of risk communication strategies while using SMPs cannot be undermined [23].

Although a tweet's engagement and sentiment can only be calculated once it has been posted, forecasting presents a fascinating way to predict the sentiments beforehand. Time series–based strategies, such as autoregressive integrated moving average (ARIMA) and vector autoregressions (VAR), have been used for forecasting emotions from SMPs [24,25]. The seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) model was recently used to gain insights into people's current emotional state via sentiment nowcasting on Twitter [26].

ML and natural language processing (NLP) algorithms have been recently used in various instances; for example, Bayesian ridge and ridge regression models were used for emotion prediction and health care analysis on large-scale data sets [27,28]. The elastic net and lasso regression have been previously used for health care access management and information exchange [29,30], while linear regression, decision tree, and random forest models are commonly used for epidemic-level disease tracking [31]. Different regression boosting algorithms, such as AdaBoost, light gradient boost , and gradient boost, have also been used for disease outbreak prediction [31]. Prophet, a Python library package, was recently used for COVID-19 outbreak prediction [32].

Objective

The implications of social media communication by HCPs have been extensively discussed [33,34]. Although they focus on the advantages and methods of extracting health- and disease-related

content from social media, there is currently a lack of understanding of how social media usage by public health agencies, nongovernment organizations (NGOs), and pharmaceutical companies resonates with society. Additionally, the study of tweets' sentiments can supplement existing models for generating content for future tweets. Predicting the tweet sentiment is 1 way to achieve this goal. Therefore, it is crucial to convert this textual content into information for formulating future strategies and gaining valuable insights into perceptions of social media users.

The remainder of the paper is structured as follows: First, a preliminary analysis of topic modeling using the best-performing clustering algorithm is presented in the Methods section, followed by sentiment and engagement analysis using CardiffNLP's *twitter-roberta-base-sentiment* model. We then conducted time series–based sentiment forecasting using 16 univariate models on the complete data set. The Results section outlines model topics obtained, which were used for generating heatmaps to obtain insights into topicwise tweets. Next, we discussed user engagement with its impact to understand whether there were specific occurrences of higher levels of engagement impacted by any offline events. In addition, we discussed results from best-performing sentiment-forecasting models. Finally, in the Discussion section, we draw conclusions and present an outline for future work.

Methods

Data Set

The data for this study (181,469 tweets) were gathered from the accounts of major US and Canadian health care organizations, pharmaceutical companies, and the World Health Organization (WHO) using the Twitter Academic API for Research v2 [35] during the time frame of January 1, 2017, to December 31, 2021. The top 5 pharmaceutical companies were selected based on the recommendations made by HCPs on Twitter [36]. Table 1 lists the number of tweets scraped for each Twitter handle. Each organization is referred to as a *user*, and the type of organization (ie, pharmaceutical company, public health agency, NGO) is referred to as a user group for the scope of this study.

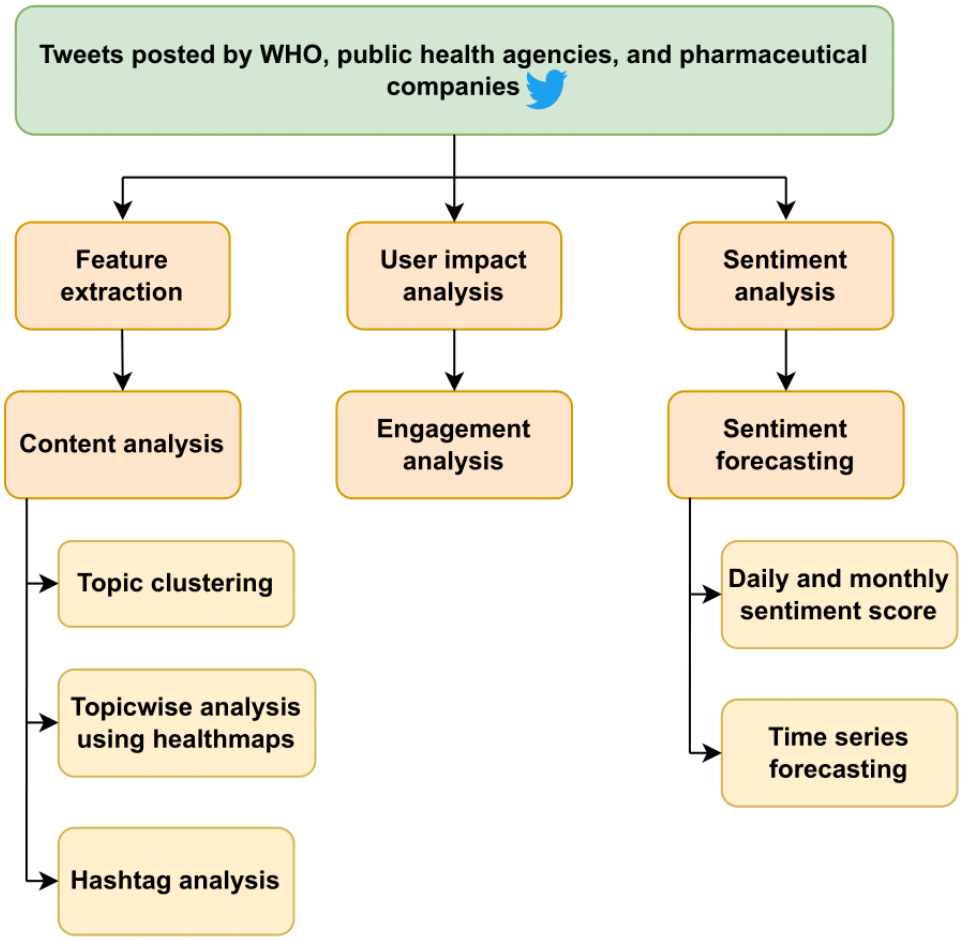
The complete timeline was divided into 2 phases for analysis, *before* COVID-19 and *during* COVID-19, based on the confirmation of the first COVID-19 community transmission case in North America on February 26, 2020 [37]. Figure 1 presents an overview of the research framework.

Table 1. Distribution of tweets for the selected user accounts of 3 types of organizations.

Name of organization (Twitter handle)	Before COVID-19, n (%)	During COVID-19, n (%)	Total tweets, N
Public health agencies			
Centers for Disease Control and Prevention (CDCgov)	8435 (58.6)	5963 (41.4)	14,398
Centers for Disease Control and Prevention (CDC_eHealth)	1376 (86.3)	219 (13.7)	1594
Government of Canada for Indigenous (GCIndigenous)	3505 (54.0)	2989 (46.0)	6494
Health Canada and PHAC (GovCanHealth)	7878 (17.2)	37,907 (82.8)	45,785
US Department of Health & Human Services (HHSGov)	7890 (56.9)	5969 (43.1)	13,859
Indian Health Service (IHSgov)	1090 (44.7)	1346 (55.3)	2436
Canadian Food Inspection Agency (InspectionCan)	4145 (62.2)	2516 (37.8)	6661
National Institutes of Health (NIH)	5837 (71.6)	2314 (28.4)	8151
National Indian Health Board (NIHB1)	1247 (51.1)	1195 (48.9)	2442
US Food and Drug Administration (US_FDA)	5810 (59.7)	3925 (40.3)	9735
Total	47,213 (42.3)	64,343 (57.7)	111,555
Pharmaceutical companies			
AstraZeneca (AstraZeneca)	3462 (78.2)	963 (21.8)	4425
Biogen (biogen)	1819 (61.9)	1120 (38.1)	2939
Glaxo SmithKline (GSK)	4200 (69.3)	1857 (30.7)	6057
Johnson & Johnson (JNJNews)	4813 (71.4)	1926 (28.6)	6739
Pfizer (pfizer)	3637 (64.1)	2039 (35.9)	5676
Total	17,931 (69.4)	7905 (30.6)	25,836
NGO[a]			
World Health Organization (WHO)	24,775 (56.2)	19,303 (43.8)	44,078

[a]NGO: nongovernment organization.

Figure 1. Overall research framework. WHO: World Health Organization.



Content Analysis

The content of each user was divided into 2 phases, before and during COVID-19. We performed topic modeling on the tweets authored by the organizations by using the topics yielded by the best-performing topic model in order to explore the most and least talked about topics with the help of heatmaps. Additionally, we examined the top 10 hashtags used by these organizations.

Preprocessing

First, all nonalphabets (numbers, punctuation, new-line characters, and extra spaces) and Uniform Resource Locators (URLs) were removed using the regular expression module (*re 2.2.1*) [38] for all tweets. The cleaned text was then tokenized using the *nltk 3.2.5* library [39]. Next, stopwords were removed, followed by stemming using PorterStemmer, and lemmatizing using the WordNetLemmatizer from *nltk*.

Topic Modeling

Researchers have used term frequency–inverse document frequency (TF-IDF) to create document embeddings for tweets [40]. Following their approach, we preprocessed and generated document embeddings for tweets and input them to 5 different clustering algorithms: LDA, parallel LDA, nonnegative matrix factorization (NMF), latent semantic indexing (LSI), and the hierarchical dirichlet process (HDP). These clustering algorithms were executed 5 times with varying random seed values. The seed values accounted for the short and noisy nature of tweets. We calculated the coherence scores of the topic models, c_{umass} [41] and c_v [42], to confirm performance consistency over multiple runs.

We used Gensim LDA [43], Gensim LDA multicore (parallel LDA) [44], and Gensim LSI [44,45] models. For NMF and HDP models, we used online NMF for large corpora [46] and online variational inference [46,47] models, respectively.

Heatmaps

Heatmaps were generated using *seaborn* to analyze the volume of tweets for each topic. The topics yielded by the best-performing topic model as per the time phase (ie, before and during COVID-19) were leveraged to generate heatmaps. Each cell represented the total count of tweets for a particular topic by an organization. For example, among pharmaceutical companies, AstraZeneca had the highest number of tweets (n=1729, 49.9%) before COVID-19 for chronic diseases.

Hashtags

The top 10 hashtags mentioned in the users' tweets were evaluated using the *advertools 0.13.0* module [48]. This tool extracts hashtags in social media posts. It was used for analyzing the similarities and differences in the tweeting behavior before and during COVID-19 and conducting topic analysis.

Sentiment Analysis

Sentiment analysis is an NLP approach used to categorize the sentiments appearing in Twitter messages based on the keywords used in each tweet. We tested different models that classify a user's tweet in 1 of 3 categories: positive, negative, and neutral. Although there is no common threshold for how many tweets should be sampled, we witnessed a range of around 2000 tweets [49-51] to several thousand tweets [52-54] when testing a model. For this study, we sampled 3000 tweets uniformly distributed over the span of our data collection time frame and from all Twitter handles. The tweets were then labeled by 3 distinct annotators, and the sentiment category with the highest votes was chosen as the overall sentiment. CardiffNLP's *twitter-roberta-base-sentiment* model [55], which is trained on a 60 million Twitter corpus, was used to obtain sentiment labels on the sampled data set. We checked for similarity between human annotations and model labels, and the similarity percentage for CardiffNLP's model was 69.96%; the model was therefore used to predict the sentiment on the remaining tweets of the users.

Engagement Analysis

For a given user, Twitter defines the engagement rate [56] as presented in Equation (1):

where "*Engagement* is the summation of the number of likes, replies, retweets, media views, tweet expansion, profile, hashtag, URL clicks, and new followers gained for every tweet, and *Impressions* is the total number of times a tweet has been seen on Twitter, such as through a follower's timeline, Twitter search, or as a result of someone liking your tweet."

Researchers have analyzed the impact (popularity) of Twitter handles by proposing heuristic and neural network–based models [57-59]. We defined it as a function of followers, following, the total number of tweets, and the profile age and calculated it using Equation (2):

where *listedCount* is the number of public lists of which this user is a member.

The total number of tweets produced by a user was considered inversely proportional to the user's impact, because a user tweeting occasionally and receiving higher engagement is more impactful than a user tweeting regularly with lower engagement.

Engagement analysis was performed to quantify the popularity of a topic generated. The engagement for each user was defined as the product of average engagement per day and their impact, as described in Equation (3). The average engagement per day was calculated as the sum of the count of likes, replies, retweets, and quotes per day. These reactions were aggregated from January 1, 2017, to December 31, 2021.

The exponential moving average (EMA) was calculated with a window span of 151 days for every user, and outliers were removed using the z-score, followed by smoothening of the average engagement per day to the eighth degree using the Savitzky-Golay filter [60].

Sentiment Forecasting

To forecast the sentiment per day, we first needed to quantify the overall sentiment of the tweets from each user every day. We leveraged CardiffNLP's *twitter-roberta-base-sentiment* model [55] to calculate the sentiments of all the tweets collected for our analysis and then calculated the daily sentiment score, as mentioned in Equation (4), based on the sentiment category with the maximum number of tweets for that day, followed by assigning the sentiment score based on the sentiment: 0 for *neutral* sentiment, the ratio of the count of positive tweets to total tweets for *positive* sentiment, and the negation of the ratio of the count of negative tweets to the total tweets for *negative* sentiment.

The daily sentiment scores were then resampled to a monthly mean sentiment score, which also helped us in handling missing values, if any. The complete timeline was divided into 2 phases (ie, before and during COVID-19), as discussed before, and the sentiment score was forecasted on 20% of the data set in each period for all user groups.

A grid search was used to find optimal hyperparameters, and 5-fold cross-validation was performed for every model. The *statsmodel* library [61] was used for ARIMA [62] and SARIMAX [63] models, and *pycaret* [64] was used for regression-based models. We also reported the performance of the *prophet* [65] model on the data set.

Three metrics, the mean absolute error (MAE), the mean square error (MSE), and the root-mean-square error (RMSE), were selected to evaluate the forecasting accuracy of the models. We considered 1-step-ahead forecasting for this study as it helped avoid problems related to cumulative errors from the preceding period.

Computational Resources

The study was performed using Compute Canada (now called the Digital Research Alliance of Canada) resources, which provide access to advanced research computing (ARC), research data management (RDM), and research software (RS). The following is a list of the computing resources offered by one of the clusters from National Services (Digital Research Alliance), Graham:

- Central processing unit (CPU): 2x Intel E5-2683 v4 Broadwell@2.1 GHz
- Memory (RAM): 30 GB

Results

Content Analysis

The details of the parameters used for each model are discussed in Multimedia Appendix 1, Table S1. Table 2 shows the mean

coherence scores (c_v and c_{umass}) for each clustering algorithm. Although the HDP had the highest c_v scores in both time phases (ie, 0.696 and 0.650 before and during COVID-19, respectively), NMF had the best c_{umass} scores (–3.653 and –3.794, respectively) and generated the most meaningful topics for the data set (see Multimedia Appendix 1, Tables S2 and S3). Therefore, the top 5 topics generated by NMF were selected to search for on the first page of Google Search results. The resulting contents were then retrieved to interpret the extracted topic keywords to propose a suitable topic name. For example, for the set of keywords yielded by the topic model "*community health*, *care*, *community health services*, *health center*, *family health centers*, *community plan*, *community clinic*, *family health care*, *qualified health centers*, *health services*," we assigned the topic *community health care*.

The scaled heatmaps showing the topic distribution for different Twitter handles are shown in Figure 2. Prior to COVID-19, chronic diseases were the most active topic, with a total of 9488 tweets from pharmaceutical companies and WHO (see Figure 2a). However, during COVID-19, we observed that COVID-19, health research, and chronic diseases were the most-discussed topics, with 52,148 tweets from all data sets combined (see Multimedia Appendix 1, Figures S1b and S1d).

This shift in the tweets' content was observed across the complete data set, and we further made the following inferences:

- Before COVID-19: Chronic diseases were the most talked about topic for pharmaceutical companies (AstraZeneca, 1729, 49.9%, tweets; Pfizer, 1168, 32.1%, tweets) and for WHO (4831, 19.5%, tweets), followed by tweets on health research (WHO, 1703, 6.9%, tweets; AstraZeneca, 1037, 29.9%, tweets). This is supported by Figure 3a, which shows #cancer, #lungcancer, #alzheimers, #hiv, and #ms to be prominently used in tweets. Among public health agencies, the NIH's and the CDC's Twitter handles were the most active, with 1840 (31.6%) and 1742 (20.6%) tweets discussing health research and chronic diseases, respectively, strongly supported by the most used hashtags #nativehealth and #foodsafety (refer to Multimedia Appendix 1, Figures S2a and S2c).
- During COVID-19: Chronic diseases and health research were the most active topics for AstraZeneca (680, 70.6%, tweets) and Glaxo SmithKline (GSK, 655, 35.2%, tweets), respectively. In addition, COVID-19 and vaccination were most talked about by GSK (398, 21.4%, tweets) and Pfizer (396, 19.4%, tweets). Figure 3b shows the hashtags supporting this: #covid19, #alzheimers, #cancer, #multiplesclerosis, and #vaccine. GovCanHealth was by far the most active public health agency on Twitter, with 16,832 (87.2%) tweets on health research, 16,449 (85.2%) tweets on vaccination, and 14,260 (73.8%) tweets on COVID-19, having #covid19, #coronavirus, and #covidvaccine as trending hashtags. The majority of the tweets by WHO were on COVID-19 (8911 tweets) and vaccination (2131 tweets), with #covid19, #coronavirus, and #vaccineequity appearing frequently in the tweets (refer to Multimedia Appendix 1, Figure S2d).

Table 2. Mean coherence scores and CPU[a] time for different clustering algorithms.

Clustering algorithm	c_v	c_{umass}	Time taken (minutes:seconds)
Before COVID-19			
LDA[b]	0.352	–5.526	17:11
Parallel LDA	0.396	–3.709	5:48
NMF[c]	0.493	–3.653	7:38
LSI[d]	0.316	–5.921	0:16
HDP[e]	0.696	–18.668	3:24
During COVID-19			
LDA	0.456	–5.688	14:01
Parallel LDA	0.446	–3.990	6:08
NMF	0.567	–3.794	7:04
LSI	0.381	–5.356	0:16
HDP	0.650	–17.610	3:01

[a]CPU: central processing unit.

[b]LDA: latent dirichlet allocation.

[c]NMF: nonnegative matrix factorization.

[d]LSI: latent semantic indexing.

[e]HDP: hierarchical dirichlet process.

Figure 2. Scaled heatmaps showing topic distribution for pharmaceutical companies before and during COVID-19.

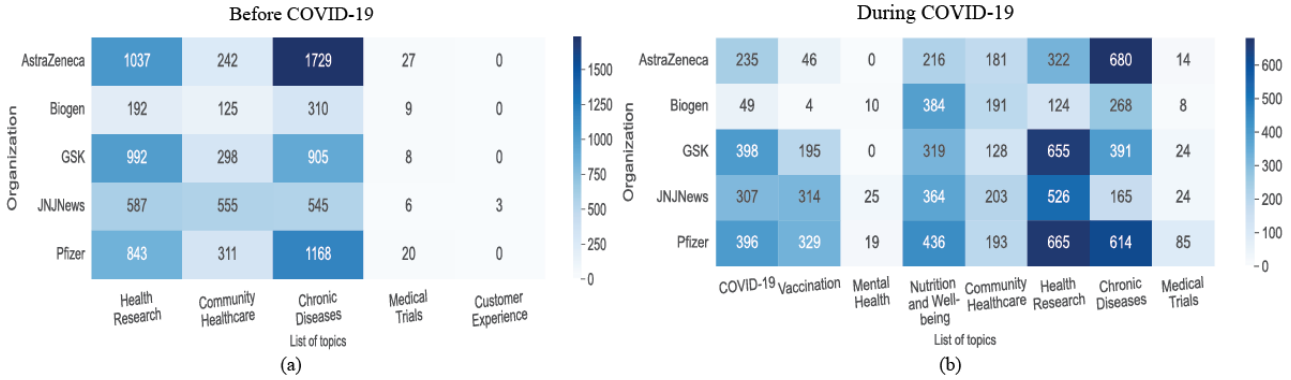
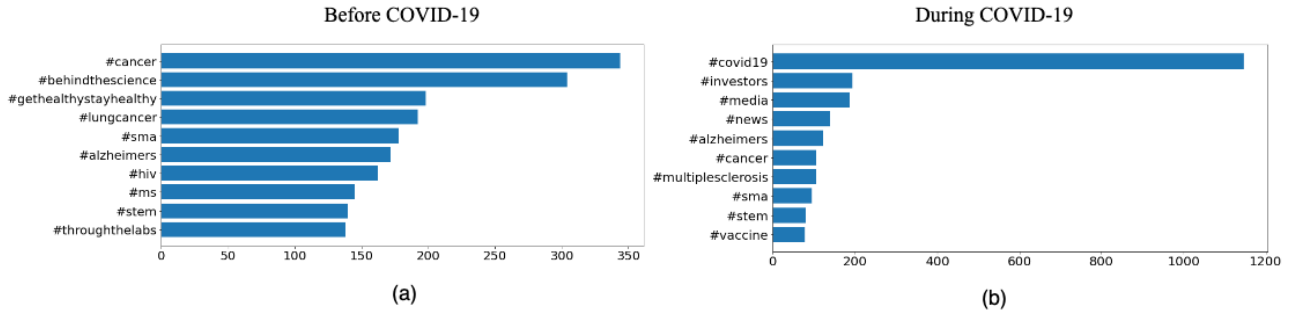


Figure 3. Top hashtags of pharmaceutical companies before and during COVID-19.



Engagement Analysis

WHO (user impact=4171.24) had the highest impact overall, followed by public health agencies (CDC user impact=2895.87; NIH user impact=891.06). Among pharmaceutical companies, Pfizer's user impact was the highest at 97.79. The user impact was normalized between the range of 0 and 1 and is shown in Figure 4.

Among pharmaceutical companies, Pfizer's user engagement was far higher than that of others (Figure 5), both before and during COVID-19, with the highest engagement observed at the time of its COVID-19 vaccine's success in November 2020. A jump in engagement was also observed in May 2021, when Pfizer announced its plan for helping India fight the second wave of coronavirus (refer to Multimedia Appendix 1, Table S4).

A similar trend was observed in public health agencies, with the CDC's account showing the highest user engagement between March and June 2020, the early months of the COVID-19 pandemic. A sharp rise in user engagement was observed in May 2021, when the CDC announced a relaxation on social distancing and masking rules for fully vaccinated individuals. The user engagement on WHO's account varied significantly over time. Its engagement was the highest in the time frame of February-April 2020, the early months of the pandemic, similar to what was observed for public health agencies. A sharp increase was seen in October 2020 following the announcement of the World Mental Health Day and in late 2020, when WHO made an announcement for COVID-19 vaccine development (refer to Multimedia Appendix 1, Figure S3).

Figure 4. User impact of all Twitter handles scaled between 0 and 1. CDC: Centers for Disease Control and Prevention; NIH: National Institutes of Health; WHO: World Health Organization.

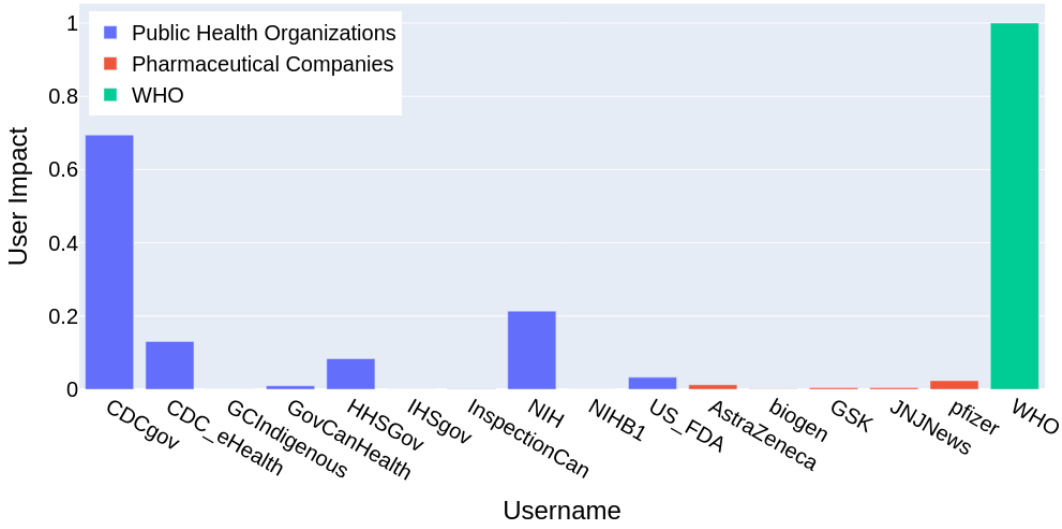
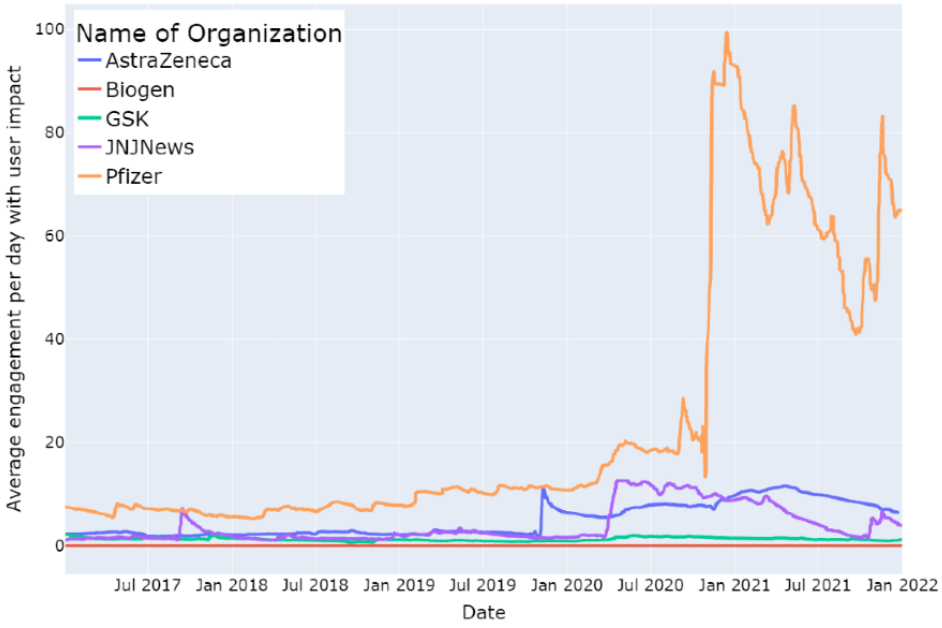


Figure 5. User engagement on Twitter accounts of pharmaceutical companies from January 1, 2017, to December 31, 2021.



Sentiment Forecasting

Table 3 shows the MAE, MSE, and RMSE for the 16 models used on the data sets. Overall, ARIMA (univariate) and SARIMAX models performed best on the majority of the subsets of the data (divided as per the organization and period), and we further made the following inferences:

- Before COVID-19: ARIMA and SARIMAX models generated the lowest MSE (0.005) and RMSE (0.072) for pharmaceutical companies. When measuring the model performance through the MAE, ARIMA performed better than all other models (0.063). A similar trend was observed for public health agencies, with ARIMA having the lowest MAE (0.027) and SARIMAX having the lowest RMSE (0.031) and a tie between them for the MSE (0.001). SARIMAX had the lowest MAE (0.054), MSE (0.004), and RMSE (0.080) on the WHO data set.
- During COVID-19: Using the CatBoost regressor gave the lowest MAE (0.072) and RMSE (0.086), while the K-neighbors regressor yielded the lowest MSE (0.008) for pharmaceutical companies. Performing regression using AdaBoost generated the lowest MAE (0.084) and RMSE (0.105) among all models used, and SARIMAX had the lowest MSE (0.011) for public health agencies. For WHO, the elastic net, lasso regression, and light gradient boosting performed equally well, with all 3 models having the same MAE (0.046) and RMSE (0.059), and SARIMAX had the lowest MSE (0.004).

Figure 6a shows the 1-step-ahead forecast for pharmaceutical companies before COVID-19 using ARIMA. The model was trained on sentiment scores from January 2017 to June 2019 and tested on data from July 2019 to February 2020 for tweets before COVID-19. The 1-step-ahead forecasting aligned well with the observed sentiment scores, and we obtained similar

results for public health agencies and WHO. The organizations showed some deviations from observed sentiments while conducting 1-step-ahead forecasting during COVID-19, making it difficult to predict their sentiment accurately, as seen in Multimedia Appendix 1, Figure S4.

To verify the forecasting performance of these models, we checked for the nature of their residual errors (ie, whether the residuals of the models were normally distributed with mean 0 and SD 1 and were uncorrelated). From Multimedia Appendix 1, Figure S5, as in the case of public health agencies, before COVID-19 using ARIMA, we confirmed the aforementioned through *plot_diagnostics*. The green kernel density estimation (KDE) line closely followed the normal distribution (N {0,1}) line in the top-right corner of Multimedia Appendix 1, Figure S5, which is a positive indicator that the residuals were scattered normally. The quantile-quantile (Q-Q) plot on the bottom left shows that the distribution of residuals (blue dots) approximately followed the linear trend of samples drawn from a standard normal distribution, N. This confirms again that the residuals were normally distributed. The residuals over time (top left in Multimedia Appendix 1, Figure S5) showed no apparent seasonality and have 0 mean. The autocorrelation plot (ie, correlogram) attested this, indicating that the time series residuals exhibited minimal correlation with lagged forms of themselves. Thus, these findings encouraged us to believe that our models provide an adequate fit, which might aid us in understanding the sentiments of the organizations and forecasting their values without overburdening our hardware with computationally heavy models.

Table 3. Results of time series sentiment forecasting using different ML[a] models (all metrics are 5-fold cross-validation).

Models	Pharmaceutical companies						Public health agencies						WHO[b]					
	Before COVID-19			During COVID-19			Before COVID-19			During COVID-19			Before COVID-19			During COVID-19		
	MAE[c]	MSE[d]	RMSE[e]	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
ARIMA[f]	0.063[g]	0.005[g]	0.072[g]	0.098	0.013	0.112	0.027[g]	0.001[g]	0.032[h]	0.240	0.082	0.286	0.066[h]	0.006[h]	0.080[h]	0.106	0.012	0.111
SARIMAX[i]	0.065[h]	0.005[g]	0.072[g]	0.084	0.011	0.104	0.028[j]	0.001[g]	0.031[g]	0.709	0.011[g]	0.106[h]	0.054[g]	0.004[g]	0.061[g]	0.047[h]	0.004[g]	0.066
Bayesian ridge	0.083	0.010	0.100	0.102	0.018	0.119	0.031	0.001	0.037	0.141	0.037	0.163	0.075[j]	0.009[j]	0.087[j]	0.061	0.008	0.075
Ridge regression	0.069	0.008	0.085	0.079	0.011	0.094	0.030	0.002	0.038	0.124	0.029	0.147	0.076	0.009	0.091	0.056	0.007	0.068
CatBoost regressor	0.066	0.007[j]	0.080[h]	0.072[g]	0.008[h]	0.086[g]	0.027[h]	0.001[h]	0.035	0.104	0.023	0.127	0.079	0.009	0.089	0.052	0.007	0.065
K-neighbors regressor	0.070	0.009	0.087	0.075[h]	0.008[g]	0.087[h]	0.030	0.001	0.036	0.093[j]	0.022	0.113	0.081	0.011	0.100	0.050	0.007	0.061[j]
Elastic net	0.070	0.008	0.088	0.080	0.009[j]	0.093[j]	0.029	0.001[h]	0.035	0.087[h]	0.021[j]	0.109[j]	0.082	0.011	0.100	0.046[g]	0.006[h]	0.059[g]
Lasso regression	0.070	0.008	0.088	0.080	0.009[j]	0.093[j]	0.029	0.001	0.035	0.087[h]	0.021[j]	0.109[j]	0.082	0.011	0.100	0.046[g]	0.006[h]	0.059[g]
Random forest regressor	0.065[j]	0.007[h]	0.081[j]	0.080	0.010	0.093	0.028	0.001[h]	0.034[j]	0.110	0.024	0.134	0.082	0.009	0.090	0.047[j]	0.006[j]	0.060[h]
Light gradient boosting machine	0.070	0.008	0.088	0.080	0.009[j]	0.093[j]	0.029	0.001[h]	0.035	0.087[h]	0.021[j]	0.109[j]	0.082	0.011	0.100	0.046[g]	0.006[h]	0.059[g]
Gradient boosting regressor	0.075	0.008	0.086	0.079	0.010	0.094	0.029	0.001[j]	0.036	0.141	0.034	0.168	0.082	0.010	0.094	0.051	0.008	0.064
AdaBoost regressor	0.070	0.007	0.082	0.080	0.010	0.091	0.029	0.001	0.037	0.084[g]	0.020[h]	0.105[g]	0.087	0.010	0.096	0.057	0.007	0.072
Extreme gradient boosting	0.068	0.009	0.087	0.080	0.011	0.098	0.031	0.002	0.040	0.151	0.045	0.171	0.087	0.011	0.098	0.055	0.007	0.065
Decision tree regressor	0.076	0.009	0.086	0.087	0.013	0.106	0.029	0.001	0.037	0.112	0.030	0.142	0.098	0.014	0.111	0.048	0.006[j]	0.061
Linear regression	0.245	0.312	0.314	0.094	0.017	0.114	0.157	0.164	0.216	0.124	0.029	0.148	2.367	52719	3.334	0.062	0.008	0.076
Prophet	0.108	0.016	0.126	0.089	0.011	0.104	0.040	0.002	0.049	0.120	0.015	0.124	0.114	0.020	0.143	0.086	0.011	0.106

[a]ML: machine learning.

[b]WHO: World Health Organization.

[c]MAE: mean absolute error.

[d]MSE: mean squared error.

[e]RMSE: root-mean-square error.

[f]ARIMA: autoregressive integrated moving average.

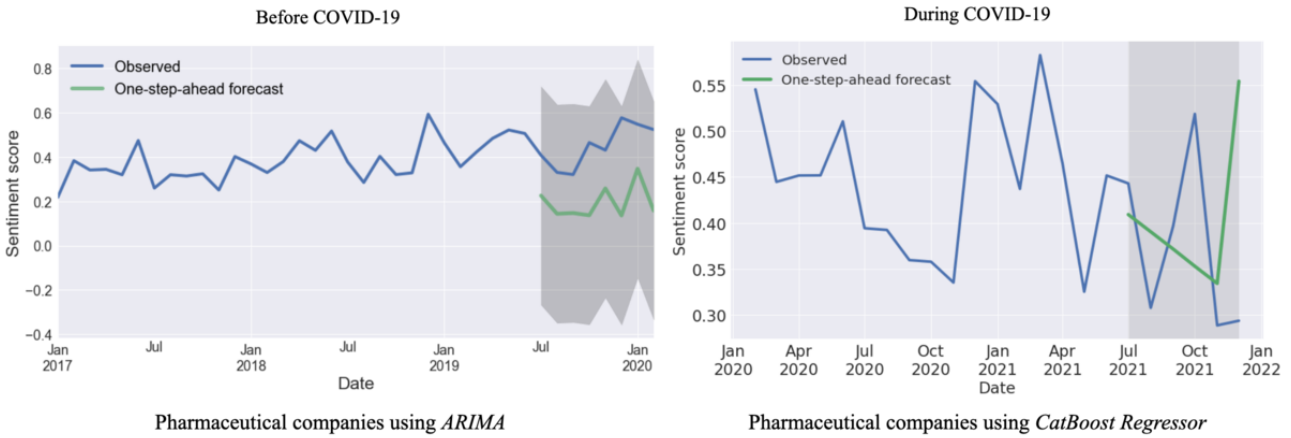
[g]The highest-performing forecasting method.

[h]The second-highest-performing forecasting method.

[i]SARIMAX: seasonal autoregressive integrated moving average with exogenous factors.

[j]The third-highest-performing forecasting method.

Figure 6. One-step-ahead forecast for all pharmaceutical companies before and during COVID-19 using the best-performing models from Table S1 (Multimedia Appendix 1). ARIMA: autoregressive integrated moving average.



Discussion

Principal Findings

In this paper, we proposed a framework for using NLP-based text-mining techniques for performing comprehensive social media content analysis of various health care organizations. We processed reasonably large amounts of textual data for topic modeling, sentiment and engagement analysis, and sentiment forecasting. Our study revealed the following key findings:

- Being the most active organization on social media does not translate to more user impact. WHO and the US public health agency CDC generated far more user impact than the Public Health Agency of Canada, even though the latter had a high number of relevant tweets when analyzed topicwise. People are more likely to engage with *neutral* tweets, which usually consist of some public health announcement rather than exclusively *positive* or *negative* tweets. This might mean that organizations can leverage this knowledge while creating content for social media posts in the future to increase their visibility in the online sphere.
- Certain topics normally translate to more user engagement. Although the content on chronic diseases and health research dominated most of the tweets posted over the study period, there was a marked shift toward a discussion on COVID-19 and vaccination for public health agencies, more than what was observed in pharmaceutical companies. Tweets on COVID-19 and chronic diseases generate more interest among the public. Perhaps surprisingly, we found that people are not much receptive to content on medical trials, often shared by pharmaceutical companies, unless it concerns a public health emergency, such as the COVID-19 pandemic. Using particular hashtags certainly helps in generating engagement, as we found that most user engagement was highly skewed toward tweets concerning COVID-19. Moreover, our study revealed that compared to the user engagement patterns found in the majority of health care organizations (ie, with peaks observed around major events or announcements), there are wide variations in user engagement for WHO. This could be due to the global presence of WHO, implying that it might not be the same set of followers engaging with its content every time, but rather only those who are impacted by or interested in the content in some way.
- When the content is structured, results tend to exceed expectations. We conducted sentiment forecasting on the data sets using different moving averages and various ML univariate models. Surprisingly, we observed that when the content is structured, as is normally the case for that available on official Twitter accounts, results tend to exceed expectations, more so before COVID-19 than during COVID-19. The models used in this research are able to predict monthwise tweet sentiment with high accuracy and low errors. This helped us in analyzing our work in-depth, and we did not need to create any multivariate ML models. Results show that commonly used ARIMA and SARIMAX models work well, and they can be used for predicting tweet sentiments on live data. This could also help organizations correlate tweet sentiment with user engagement. For example, the highest engagement on Pfizer's tweets was for the ones labeled *neutral*, implying that the organization should structure the content of its future tweets in a similar manner to maintain higher levels of engagement. Furthermore, tweets that mention more news-relevant content might be able to translate it into more user engagement.

Limitations and Future Work

There are 3 limitations of this study that could be addressed in future research. First, this work focused on dividing the tweets into 2 phases, *before* and *during* COVID-19. In the future, researchers can pursue other methods of structuring the analysis timeline. Second, this study dealt with only the structured textual content of tweets. It would be interesting to also incorporate the presence of image attributes in future studies. Finally, as the scope of this study was limited to health care organizations, we did not account for public demographics. Understanding the demographic background of the public engaging with this content is another area that can be explored in future studies.

Conclusion

This study examined the online activity of US and Canadian health care organizations on Twitter. The NLP-based analysis of social media presented here can be incorporated to gauge

engagement on the previously published tweets and to generate tweets that create an impact on people accessing health information via SMPs. As organizations continue to leverage SMPs by providing the latest information to the community, predicting a tweet's sentiment before publishing can boost an organization's perception by the public. In conclusion, we found that performing content analysis and sentiment forecasting on an organization's social media usage provides a comprehensive view of how it resonates with society.

Acknowledgments

The authors thank members of the DaTALab at Lakehead University for valuable discussions, along with Andy Pan, Chandreen Ravihari Liyanage, and Lakshmi Preethi Kamak for annotating the sampled tweets to evaluate the tweet sentiment. This study was conducted using Digital Research Alliance of Canada computing resources. AS and MKB were supported by Vector Scholarships in artificial intelligence (AI) from Vector Institute, Toronto, Canada, and a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (#RGPIN-2017-05377) held by VM.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Topics and user engagement.
[PDF File (Adobe PDF File), 666 KB - medinform_v10i8e37829_app1.pdf]

References

1. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014 Jul;39(7):491-520 [FREE Full text] [Medline: 25083128]
2. Househ M. The use of social media in healthcare: organizational, clinical, and patient perspectives. Stud Health Technol Inform 2013;183:244-248. [Medline: 23388291]
3. Zhou L, Zhang D, Yang CC, Wang Y. Harnessing social media for health information management. Electron Commer Res Appl 2018 Jan;27:139-151 [FREE Full text] [doi: 10.1016/j.elerap.2017.12.003] [Medline: 30147636]
4. Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, et al. Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach. J Med Internet Res 2020 Nov 25;22(11):e20550 [FREE Full text] [doi: 10.2196/20550] [Medline: 33119535]
5. Benetoli A, Chen T, Aslani P. How patients' use of social media impacts their interactions with healthcare professionals. Patient Educ Couns 2018 Mar;101(3):439-444. [doi: 10.1016/j.pec.2017.08.015] [Medline: 28882545]
6. Li H, Sakamoto Y. Social impacts in social media: an examination of perceived truthfulness and sharing of information. Comput Hum Behav 2014 Dec;41:278-287. [doi: 10.1016/j.chb.2014.08.009]
7. Lu Y, Wu Y, Liu J, Li J, Zhang P. Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. J Med Internet Res 2017 Apr 07;19(4):e109 [FREE Full text] [doi: 10.2196/jmir.7087] [Medline: 28389418]
8. Tyrawski J, DeAndrea DC. Pharmaceutical companies and their drugs on social media: a content analysis of drug information on popular social media sites. J Med Internet Res 2015 Jun 01;17(6):e130 [FREE Full text] [doi: 10.2196/jmir.4357] [Medline: 26032738]
9. Abualigah L, Alfar H, Shehab M. Sentiment analysis in healthcare: a brief review. In: Abd Elaziz M, Al-qaness MAA, Ewees AA, editors. Recent Advances in NLP: The Case of Arabic Language. Cham: Springer International; 2020:129-141.
10. Chandrasekaran R, Mehta V, Valkunde T, Moustakas E. Topics, trends, and sentiments of tweets about the COVID-19 pandemic: temporal infoveillance study. J Med Internet Res 2020 Oct 23;22(10):e22624 [FREE Full text] [doi: 10.2196/22624] [Medline: 33006937]
11. Poddar S, Mondal M, Misra J. Winds of Change: Impact of COVID-19 on Vaccine-Related Opinions of Twitter Users. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/19334 [accessed 2022-06-29]
12. Rufai S, Bunce C. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. J Public Health (Oxf) 2020 Aug 18;42(3):510-516 [FREE Full text] [doi: 10.1093/pubmed/fdaa049] [Medline: 32309854]
13. Haman M. The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic. Heliyon 2020 Nov;6(11):e05540 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e05540] [Medline: 33294685]
14. Rosenberg H, Syed S, Rezaie S. The Twitter pandemic: the critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. CJEM 2020 Jul 06;22(4):418-421 [FREE Full text] [doi: 10.1017/cem.2020.361] [Medline: 32248871]
15. Park HW, Park S, Chong M. Conversations and medical news frames on Twitter: infodemiological study on COVID-19 in South Korea. J Med Internet Res 2020 May 05;22(5):e18897 [FREE Full text] [doi: 10.2196/18897] [Medline: 32325426]

16. Hussain A, Tahir A, Hussain Z, Sheikh Z, Gogate M, Dashtipour K, et al. Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. J Med Internet Res 2021 Apr 05;23(4):e26627 [FREE Full text] [doi: 10.2196/26627] [Medline: 33724919]
17. Lwin MO, Lu J, Sheldenkar A, Schulz PJ, Shin W, Gupta R, et al. Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. JMIR Public Health Surveill 2020 May 22;6(2):e19447 [FREE Full text] [doi: 10.2196/19447] [Medline: 32412418]
18. Dubey AD. Twitter sentiment analysis during COVID19 Outbreak. SSRN Electron J 2020:1-9. [doi: 10.2139/ssrn.3572023]
19. Gao S, He L, Chen Y, Li D, Lai K. Public perception of artificial intelligence in medical care: content analysis of social media. J Med Internet Res 2020 Jul 13;22(7):e16649 [FREE Full text] [doi: 10.2196/16649] [Medline: 32673231]
20. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on Twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res 2021 Feb 10;23(2):e25431 [FREE Full text] [doi: 10.2196/25431] [Medline: 33497352]
21. Tang L, Liu W, Thomas B, Tran HTN, Zou W, Zhang X, et al. Texas public agencies' tweets and public engagement during the COVID-19 pandemic: natural language processing approach. JMIR Public Health Surveill 2021 Apr 26;7(4):e26720 [FREE Full text] [doi: 10.2196/26720] [Medline: 33847587]
22. Koumpouros Y, Toulias TL, Koumpouros N. The importance of patient engagement and the use of social media marketing in healthcare. Technol Health Care 2015 Jul 21;23(4):495-507. [doi: 10.3233/thc-150918]
23. Slavik CE, Buttle C, Sturrock SL, Darlington JC, Yiannakoulias N. Examining tweet content and engagement of Canadian public health agencies and decision makers during COVID-19: mixed methods analysis. J Med Internet Res 2021 Mar 11;23(3):e24883 [FREE Full text] [doi: 10.2196/24883] [Medline: 33651705]
24. Tommasel A, Diaz-Pace A, Rodriguez JM, Godoy D. Forecasting mental health and emotions based on social media expressions during the COVID-19 pandemic. Inf Discov Deliv 2021 Jun 03;49(3):259-268. [doi: 10.1108/idd-01-2021-0003]
25. McClellan C, Ali MM, Mutter R, Kroutil L, Landwehr J. Using social media to monitor mental health discussions - evidence from Twitter. J Am Med Inform Assoc 2017 May 01;24(3):496-502 [FREE Full text] [doi: 10.1093/jamia/ocw133] [Medline: 27707822]
26. Miliou I, Pavlopoulos J, Papapetrou P. Sentiment nowcasting during the COVID-19 pandemic. In: Discovery Science. Cham: Springer International; 2021:218-228.
27. Harper R, Southern J. A Bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. IEEE Trans Affective Comput 2022 Apr 1;13(2):985-991 [FREE Full text] [doi: 10.1109/TAFFC.2020.2981610]
28. Deepa N, Prabadevi B, Maddikunta PK, Gadekallu TR, Baker T, Khan MA, et al. An AI-based intelligent system for healthcare analysis using Ridge-Adaline stochastic gradient descent classifier. J Supercomput 2020 May 30;77(2):1998-2017. [doi: 10.1007/s11227-020-03347-2]
29. Barrera Ferro D, Brailsford S, Bravo C, Smith H. Improving healthcare access management by predicting patient no-show behaviour. Decis Support Syst 2020 Nov;138:113398. [doi: 10.1016/j.dss.2020.113398]
30. Li Y, Vinzamuri B, Reddy CK. Constrained elastic net based knowledge transfer for healthcare information exchange. Data Min Knowl Disc 2014 Dec 23;29(4):1094-1112. [doi: 10.1007/s10618-014-0389-3]
31. Singh R, Singh R. Applications of sentiment analysis and machine learning techniques in disease outbreak prediction – A review. Mater Today 2021 May:1-6 [FREE Full text] [doi: 10.1016/j.matpr.2021.04.356]
32. Mengistie T. COVID-19 outbreak data analysis and prediction modeling using data mining technique. Int J Comput 2020;38:37-60 [FREE Full text]
33. Denecke K, Nejdl W. How valuable is medical social media data? Content analysis of the medical web. Inf Sci 2009 May 30;179(12):1870-1880. [doi: 10.1016/j.ins.2009.01.025]
34. Nawaz MS, Bilal M, Lali MI, Ul Mustafa R, Aslam W, Jajja S. Effectiveness of social media data in healthcare communication. J Med Imaging Health Inform 2017 Oct 01;7(6):1365-1371. [doi: 10.1166/jmihi.2017.2148]
35. Twitter API: Academic Research Access. URL: https://developer.twitter.com/en/products/twitter-api/academic-research [accessed 2022-07-05]
36. Kangley M. HCPs Discuss 'Booster Shot' to Decrease the High Spread of the Delta Variant. URL: https://creation.co/knowledge/hcps-discuss-booster-shot-to-decrease-the-high-spread-of-the-delta-variant/ [accessed 2022-07-05]
37. CDC COVID-19 Response Team, Jorden MA, Rudman SL, Villarino E, Hoferka S, Patel MT, et al. Evidence for limited early spread of COVID-19 within the United States, January-February 2020. Morb Mortal Wkly Rep 2020 Jun 05;69(22):680-684 [FREE Full text] [doi: 10.15585/mmwr.mm6922e1] [Medline: 32497028]
38. PyPI. regex 2022.7.9. URL: https://pypi.org/project/regex/ [accessed 2022-07-05]
39. PyPI. nltk 3.7. URL: https://pypi.org/project/nltk/ [accessed 2022-07-05]
40. Lilleberg J, Zhu Y, Zhang Y. Support vector machines and Word2vec for text classification with semantic features. 2015 Presented at: IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC); July 6-8, 2015; Beijing, China. [doi: 10.1109/icci-cc.2015.7259377]
41. Newman D, Lau J, Grieser K. Automatic evaluation of topic coherence. 2010 Presented at: Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics; June 2-4, 2010; Los Angeles URL: https://aclanthology.org/N10-1012

42. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. 2015 Presented at: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; 2015; New York, NY. [doi: 10.1145/2684822.2685324]
43. Gemsim. Latent Dirichlet Allocation. URL: https://radimrehurek.com/gensim/models/ldamodel.html [accessed 2022-07-05]
44. Gensim. Parallelized Latent Dirichlet Allocation. URL: https://radimrehurek.com/gensim/models/ldamulticore.html [accessed 2022-07-05]
45. Gensim. Latent Semantic Indexing. URL: https://radimrehurek.com/gensim/models/lsimodel.html [accessed 2022-07-05]
46. Gensim. Non-Negative Matrix Factorization. URL: https://radimrehurek.com/gensim/models/nmf.html [accessed 2022-07-05]
47. Gensim. Hierarchical Dirichlet Process. URL: https://radimrehurek.com/gensim/models/hdpmodel.html [accessed 2022-07-05]
48. PyPI. advertools 0.13.1. URL: https://pypi.org/project/advertools/ [accessed 2022-07-05]
49. Alomari K, ElSherif H, Shaalan K. Arabic tweets sentimental analysis using machine learning. In: Advances in Artificial Intelligence: From Theory to Practice. Cham: Springer International; 2017:602-610.
50. Peisenieks J, Skadins R. Uses of machine translation in the sentiment analysis of tweets. 2014 Presented at: Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference; 2014; Kaunas, Lithuania p. 2014. [doi: 10.3233/978-1-61499-442-8-126]
51. Şaşmaz E, Tek F. Tweet sentiment analysis for cryptocurrencies. 2021 Presented at: 6th International Conference on Computer Science and Engineering (UBMK); September 15-17, 2021; Ankara, Turkey p. 613-618. [doi: 10.1109/ubmk52708.2021.9558914]
52. Golubev A, Loukachevitch N. Improving results on Russian sentiment datasets. In: Communications in Computer and Information Science. Cham: Springer International; 2020:109-121.
53. Nabil M, Aly M, Atiya A. ASTD: Arabic Sentiment Tweets Dataset. 2015 Presented at: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015; Lisbon, Portugal. [doi: 10.18653/v1/d15-1299]
54. Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS One 2021 Feb 25;16(2):e0245909 [FREE Full text] [doi: 10.1371/journal.pone.0245909] [Medline: 33630869]
55. Hugging Face. cardiffnlp / twitter-roberta-base-sentiment. URL: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment [accessed 2022-07-19]
56. About Your Activity Dashboard. URL: https://help.twitter.com/en/managing-your-account/using-the-tweet-activity-dashboard [accessed 2022-07-05]
57. Daniluk M, Dabrowski J, Rychalska B. Synerise at RecSys 2021: Twitter user engagement prediction with a fast neural model. 2021 Presented at: RecSysChallenge '21: Proceedings of the Recommender Systems Challenge 2021; 2021; New York, NY. [doi: 10.1145/3487572.3487599]
58. Razis G, Anagnostopoulos I. InfluenceTracker: rating the impact of a Twitter account. 2014 Presented at: IFIP International Conference on Artificial Intelligence Applications and Innovations; September 19-21, 2014; Rhodes, Greece. [doi: 10.1007/978-3-662-44722-2_20]
59. Son J, Lee J, Oh O, Lee HK, Woo J. Using a heuristic-systematic model to assess the Twitter user profile's impact on disaster tweet credibility. Int J Inf Manag 2020 Oct;54:102176. [doi: 10.1016/j.ijinfomgt.2020.102176]
60. Marinai S, Dengel A. Document Analysis Systems VI: 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004, Proceedings. Berlin, Heidelberg: Springer; 2004.
61. statsmodels. URL: https://www.statsmodels.org/stable/index.html [accessed 2022-07-05]
62. statsmodels.tsa.arima.model.ARIMA. URL: https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima.model.ARIMA.html [accessed 2022-07-05]
63. statsmodels.tsa.statespace.sarimax.SARIMAX. URL: https://www.statsmodels.org/devel/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html [accessed 2022-07-05]
64. PyPI. pycaret. URL: https://pypi.org/project/pycaret/ [accessed 2022-07-05]
65. PyPI. prophet. URL: https://pypi.org/project/prophet/ [accessed 2022-07-05]

Abbreviations

ARC: advanced research computing
ARIMA: autoregressive integrated moving average
CDC: Centers for Disease Control and Prevention
CPU: central processing unit
HCP: health care professional
HDP: hierarchical dirichlet process
LDA: latent dirichlet allocation
LSI: latent semantic indexing
MAE: mean absolute error
ML: machine learning

MSE: mean squared error
NGO: nongovernment organization
NIH: National Institutes of Health
NLP: natural language processing
NMF: nonnegative matrix factorization
RMSE: root-mean-square error
SARIMAX: seasonal autoregressive integrated moving average with exogenous factors
SMP: social media platform
TF-IDF: term frequency–inverse document frequency
WHO: World Health Organization

Edited by T Hao; submitted 09.03.22; peer-reviewed by S Doan, A Benis; comments to author 27.06.22; revised version received 08.07.22; accepted 15.07.22; published 18.08.22.

Please cite as:
Singhal A, Baxi MK, Mago V
Synergy Between Public and Private Health Care Organizations During COVID-19 on Twitter: Sentiment and Engagement Analysis Using Forecasting Models
JMIR Med Inform 2022;10(8):e37829
URL: https://medinform.jmir.org/2022/8/e37829
doi:10.2196/37829
PMID:35849795

Original Paper

Perceptions and Discussions of Snus on Twitter: Observational Study

Jiarui Chen[1]; Siyu Xue[1], BSc; Zidian Xie[2], PhD; Dongmei Li[2], PhD

[1]Goergen Institute for Data Science, University of Rochester, Rochester, NY, United States

[2]Department of Clinical & Translational Research, University of Rochester Medical Center, Rochester, NY, United States

Corresponding Author:
Dongmei Li, PhD
Department of Clinical & Translational Research
University of Rochester Medical Center
265 Crittenden Boulevard CU 420708
Rochester, NY, 14642-0708
United States
Phone: 1 585 276 7285
Email: Dongmei_Li@urmc.rochester.edu

Abstract

Background: With the increasing popularity of snus, it is essential to understand the public perception of this oral tobacco product. Twitter—a popular social media platform that is being used to share personal experiences and opinions—provides an ideal data source for studying the public perception of snus.

Objective: This study aims to examine public perceptions and discussions of snus on Twitter.

Methods: Twitter posts (tweets) about snus were collected through the Twitter streaming application programming interface from March 11, 2021, to February 26, 2022. A temporal analysis was conducted to examine the change in number of snus-related tweets over time. A sentiment analysis was conducted to examine the sentiments of snus-related tweets. Topic modeling was applied to tweets to determine popular topics. Finally, a keyword search and hand-coding were used to understand the health symptoms mentioned in snus-related tweets.

Results: The sentiment analysis showed that the proportion of snus-related tweets with a positive sentiment was significantly higher than the proportion of negative sentiment tweets (4341/11,631, 37.32% vs 3094/11,631, 26.60%; $P<.001$). The topic modeling analysis revealed that positive tweets focused on snus's harm reduction and snus use being an alternative to smoking, while negative tweets focused on health concerns related to snus. Mouth and respiratory symptoms were the most mentioned health symptoms in snus-related tweets.

Conclusions: This study examined the public perception of snus and popular snus-related topics discussed on Twitter, thus providing a guide for policy makers with regard to the future formulation and adjustment of tobacco regulation policies.

(JMIR Med Inform 2022;10(8):e38174) doi:10.2196/38174

KEYWORDS

snus; Twitter; sentiment; topic modeling; smokeless tobacco products

Introduction

Smokeless tobacco is a type of tobacco that is neither smoked nor burnt during consumption. Examples of smokeless tobacco products include chewing tobacco, dissolvable tobacco, and oral nicotine pouches. According to the Centers for Disease Control and Prevention (CDC), in 2020, there were 5.7 million adult users of smokeless tobacco nationwide in the United States [1]. Among the smokeless tobacco products, snus is a smokeless and sometimes flavored tobacco product for oral consumption that originated from Sweden. It is usually in the following two forms: loose ground powder and sachets. When snus is consumed, it is held behind the upper lip [2]. Although this tobacco product was banned in the member countries of the European Union, with a few exceptions such as Sweden [3], its use in the rest of the world is prevalent. By 2013 for example, 18% of adolescents had tried snus in Finland [4]. In the United States, a study conducted in 2021 by the CDC suggested that 1.2% of US high school students are current users of smokeless products, including snus [1].

Studies have found that snus use may result in oral cancer, cardiovascular diseases, respiratory diseases, diabetes, and other illnesses [5]. A cohort study on 135,036 male, Swedish construction industry employees found that the age-adjusted relative risk of dying from cardiovascular disease for smokeless tobacco users was 40% higher than that for nonusers [6]. Despite these concerns, previous studies indicated that snus use has a considerably lower health risk than cigarette smoking [2,7]. According to a review on multiple health symptoms, including oral health and cardiovascular diseases, among others, the health risk of snus is significantly lower than that of cigarettes [2].

Similar to other tobacco products, snus use results in nicotine dependence, and the perceptions toward the relationship between snus consumption and other types of nicotine consumption have been controversial [5]. The gateway hypothesis states that the use of snus may lead to more addictive smoking behaviors. On the contrary, the pathway hypothesis claims that snus use helps to prevent people from smoking [5]. Previous studies on this topic reported different conclusions. A previous study tracked 496 pairs of users and nonusers of smokeless tobacco products and concluded that there was insufficient evidence to conclude that using smokeless tobacco products leads to a higher chance of smoking [8]. Another research study on smokers in Sweden found that 76.3% of the male smokers and 71.6% of the female smokers included in the study quit smoking after they started consuming snus [9]. However, a focus group study that was performed on 66 participants in 2010 concluded that the participants believed that snus use could potentially lead to smoking [10].

With the controversial gateway and pathway hypotheses and the potential health impact of snus products, disagreements on the perception of snus product may exist among the public. As snus is becoming increasingly popular, governmental regulation plays an essential role in the relationship between snus consumption and public health. For example, the US Food and Drug Administration stipulates that for smokeless tobacco products, including snus, special warnings such as "WARNING: Smokeless tobacco is addictive" should be attached to the packages [11]. For governors and regulators to better manage the relationship between snus and public health and be more informed in policy making, it is beneficial to understand how the public truly perceives snus.

Twitter, as a popular social media platform, has been used to examine smoking behaviors and perceptions of tobacco products, such as e-cigarettes [12,13]. Although perceptions of snus have been investigated by using focus groups, the sample sizes of such focus groups are very limited [10,14]. Research that uses social media data to study the public perceptions of snus is scarce.

Our study aimed to examine the public perceptions of and popular topics regarding snus on Twitter. Our study consisted of 3 specific goals. First, we aimed to determine the sentiments of snus-related tweets via a sentiment analysis. Second, we attempted to explore specific topics related to snus. Finally, we tried to examine potential health risks that were mentioned in snus-related tweets. Through a comprehensive examination of the public perceptions and the top topics discussed about snus, we hope to provide some insights to policy makers on regulating snus for public health protection.

Methods

Ethics Approval

We only used publicly available tweets for this study, and there was no identifying information on Twitter users in this study. In addition, this study was reviewed and approved by the Office for Human Subject Protection Research Subjects Review at the University of Rochester (study ID: STUDY00006570).

Data Collection and Preprocessing

We collected Twitter posts (tweets) related to snus from March 11, 2021, to February 26, 2022, through the Twitter streaming application programming interface by using the keyword *snus*, and we obtained a data set with 28,427 tweets. We then preprocessed the data to enhance their quality. First, all the tweets were lowercased. Afterward, by using the Regular Expression Operations Package (Python Software Foundation) [15], we removed the parts of tweets that did not contribute to the tweets' actual contents, including email addresses, new-line characters, single quotation marks, URLs, and "@" signs (used to mention other users). Next, we applied 2 sets of promotion filters to eliminate tweets that were related to the commercial promotion of snus [13]. The first filter targeted the usernames, using keywords such as *snus*, *smokeless*, *dealer*, *supply*, *nicotine*, *cigarette*, and *store*. Tweets posted by users with usernames containing any of these words were not included in this study because they might have been posted by commercial accounts. The second layer of the filter aimed to remove potentially commercial tweet content, and the keywords included *order*, *new*, *offer*, *discount*, and *free shipping*. Tweets that contained these words were highly likely to be promotional tweets. Finally, we eliminated the repetitive tweets. After preprocessing, the final data set contained 11,631 tweets.

Sentiment Analysis

Sentiment analysis is a computational method of learning the attitudes in text, and the Valence Aware Dictionary and Sentiment Reasoner (VADER) is a sentiment analysis package that is specialized for social media data [16]. By applying the VADER on each tweet, we assigned each tweet a sentiment score of between −1.0 and 1.0. To better define the sentiments, we grouped the tweets into 3 categories based on the corresponding sentiment scores; tweets with a sentiment score of ≥0.05 were labeled as "positive," and tweets with a score of ≤−0.05 were labeled as "negative." The remaining tweets were labeled as "neutral." The proportions of positive, neutral, and negative tweets were then calculated. The daily proportion of positive tweets was then calculated.

We performed the chi-square goodness-of-fit test by using statistical analysis software (R version 4.0.2; R Foundation for Statistical Computing) to examine the frequency distribution of different attitudes [17]. A significance level of .05 was used to determine whether the proportion of positive tweets was statistically significantly higher than the proportion of the negative tweets.

Topic Modeling

Topic modeling is a computational method of identifying major topics in text. The model we chose for our study was the latent Dirichlet allocation model, which was applied to positive tweets, neutral tweets, and negative tweets to observe the main topics that Twitter users had been discussing.

By using the *gensim* package in Python [18], we built a bigram and trigram based on our data set. Bigrams and trigrams are sequences of 2 words and 3 words, respectively. With the bigram and the trigram, we treated some of the most frequently mentioned phrases as a whole instead of 2 or 3 separate words. For example, *harm reduction* was a frequently mentioned phrase among the tweets, and we considered *harm reduction* as a single token that contributed to a topic instead of preserving *harm* and *reduction* separately.

We applied the Natural Language Toolkit to remove the stop words in the tweets [19]. Stop words include but are not limited to commonly used articles, pronouns, and propositions, which undermine the quality of topic modeling results if kept. In addition, we used spaCy (Explore) to lemmatize the words in tweets into their dictionary forms without changing their meaning [20]. For example, *smoked* became *smoke* after lemmatization. After conversion, words like *smoked* were left unused for topic modeling, and only their dictionary forms were included. Both coherence scores and intertopic distance maps were used to determine the optimal number of topics discussed in the tweets, using the *pyLDAvis* package in Python [21].

To better interpret the results from the model, we inferred the topics based on the keyword outputs and example tweets. Two authors reviewed the tweets from each category and summarized the topics independently. The results from the two authors were compared and discussed. Any discrepancy was resolved by a group of 4 members.

Health-Related Discussion

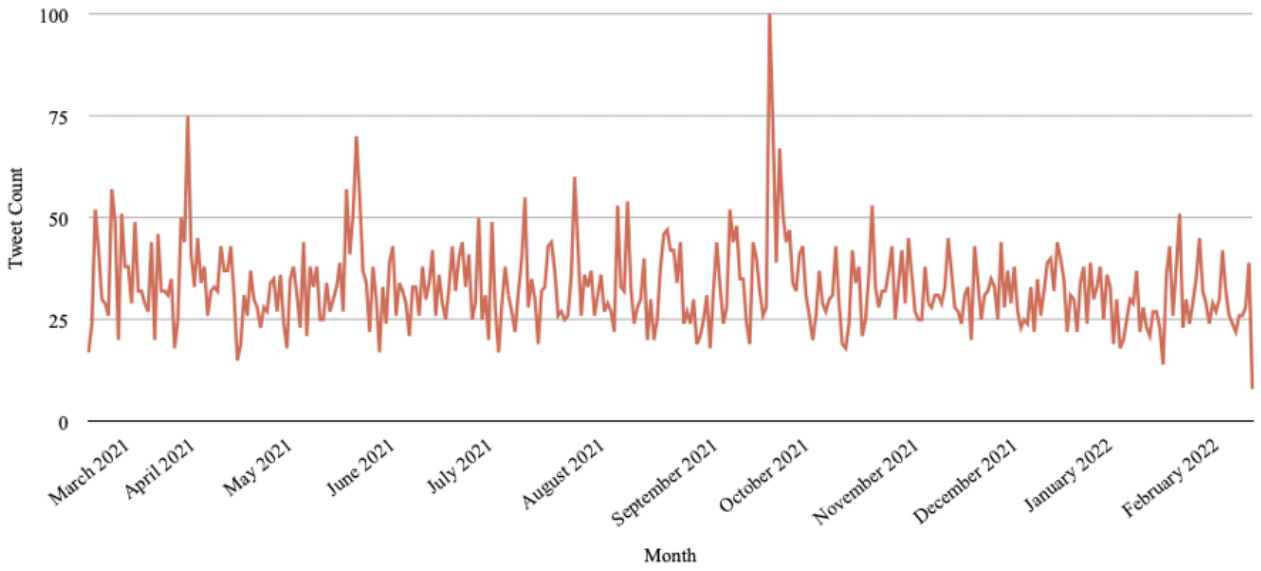
To determine the frequency of health effects that were mentioned in snus-related tweets, we filtered the data set by using a list of health-related keywords that were created in previous studies [22-24], which resulted in a set of 654 unique tweets with 1254 health-related keyword appearances. The list included the following nine major groups of health effects that are related to smoking and nicotine consumption: mouth (eg, gum, teeth, etc), respiratory (eg, lung, cough, etc), cardiovascular (eg, heart, etc), psychological (eg, stress, anxiety, etc), neurological (eg, numb, fatigue, etc), cancer (eg, lung cancer, mouth cancer, etc), throat, digestive, and other effects (eg, skin, liver, etc). For each major group of health effects, the number of occurrences of specific keywords belonging to the groups were counted. In addition, two authors hand-coded 200 randomly selected tweets to determine whether the users directly experienced the health symptoms mentioned or whether they believed that snus use might help with lowering the risk of the symptoms when compared to smoking. The Cohen κ statistic reached 0.73, indicating substantial agreement between the two coders.

Results

Temporal Analysis

To better understand the popularity of snus discussion, we examined the number of snus-related tweets over time during our study time period. As shown in Figure 1, the number of tweets per day typically oscillated between 25 and 50, with a few peaks occurring in April 10, 2021; May 31, 2021; and October 3, 2021.

Figure 1. Snus-related tweets from March 11, 2021, to February 26, 2022.



Perceptions of Snus on Twitter

To examine the public perception of snus on Twitter, we performed a sentiment analysis on tweets related to snus. The average sentiment score for 11,631 snus-related tweets was 0.080, which indicated that the overall sentiment in snus-related tweets was positive. Among these tweets, there were 4341 (37.32%) positive tweets, 3094 (26.60%) negative tweets, and 4196 (36.08%) neutral tweets. Further statistical analysis showed that the proportion of positive tweets was significantly higher than the proportion of negative tweets (4341/11,631, 37.32% vs 3094/11,631, 26.60%; $P<.001$). Our longitudinal analysis showed that there was no noticeable change in the proportion of positive tweets over time (Multimedia Appendix 1).

Topics Discussed in Snus-Related Tweets

To understand what might be responsible for different sentiments in snus-related tweets, we performed topic modeling for the tweets in the different sentiment groups. As shown in Table 1, among the positive sentiment snus-related tweets, the most popular topic was "Snus being a safer way of nicotine consumption" (1472/4341, 33.9%), followed by "Way of snus consumption" (1441/4341, 33.2%) and "Snus addiction and enjoyment" (1428/4341, 32.9%). Among the negative sentiment snus-related tweets, the top topic was "Risk comparison between snus and smoking" (1064/3094, 34.4%), followed by "Negative health impacts" (1018/3094, 32.9%) and "Other problems related to snus" (1012/3094, 32.7%). The topics for neutral sentiment snus-related tweets are included in Table S1 in Multimedia Appendix 2.

Table 1. Topics discussed in snus-related tweets with different sentiments.

Sentiment group and inferred topic	Keywords	Token percentage	Examples
Positive			
Snus addiction and enjoyment	*snus*, *good*, *make*, *time*, *day*, *love*, *feel*, *free*, *access*, *today*, *strong*, *man*, *back*, *coffee*, *pack*, *life*, *pretty*, *friend*, *enjoy*, and *week*	32.9	"Proper pint of bitter and a wintergreen snus. Perfect on a fair night like tonight"
Snus being a safer way of nicotine consumption	*pouch*, *vape*, *smoking*, *smoke*, *quit*, *cigarette*, *nicotine*, *safe*, *give*, *amp*, *year*, *alternative*, *smoker*, *start*, *risk*, *big*, *stop*, *switch*, *low*, and *option*	33.9	"For long-term nicotine use, data on safety are strongest for snus: decades of epidemiological studies. No harm. So if many people with mental health issues self-medicate with #safernicotine (they are), at least there is no harm. #qualityoflife"
Way of snus consumption	*snus*, *tobacco*, *product*, *Swedish*, *people*, *chew*, *work*, *thing*, *great*, *smokeless*, *put*, *dip*, *find*, *call*, *play*, *gum*, *nice*, *hard*, *flavor*, and *mouth*	33.2	"snus is a black tobacco product you chew or put on your gums. You don't snort it or sniff it. He's clearly closing one nostril to sniff smelling salts, which are commonly used in sports. Not rocket science."
Negative			
Risk comparison between snus and smoking	*Tobacco*, *smoke*, *vape*, *cigarette*, *smoking*, *pouch*, *product*, *cancer*, *risk*, *low*, *nicotine*, *amp*, *quit*, *harm*, *gum*, *rate*, *smoker*, *chew*, *reduce*, and *smokeless*	34.4	"not just snus but the attempt to restrict and eliminate all lower risk products is astonishingly short sighted."
Negative health impacts	*Snus*, *ban*, *make*, *stop*, *day*, *Swedish*, *year*, *give*, *feel*, *thing*, *death*, *man*, *start*, *high*, *mouth*, *kill*, *lose*, *addiction*, *long*, and *cig*	32.9	"not in epok which i assume is some zoomer snus? i dont know i only use odens and sometimes siberia which has no flavouring just tobacco. the nicotine content is pretty potent in those, would kill your average vaper no joke."
Other problems related to snus	*snus*, *people*, *time*, *bad*, *work*, *put*, *hard*, *good*, *study*, *week*, *today*, *back*, *call*, *big*, *find*, *coffee*, *problem*, and *life*	32.7	"our big daddy is always the leader he is the familys captain and chief, but once i choked when my snus caught up in my throat cause there was our pop in the oak."

Health Risks Mentioned in Snus-Related Tweets

To understand what health risks might be associated with snus, we explored the health symptoms mentioned in the snus-related tweets. Oral health (mouth effects) was the most mentioned health category in snus-related tweets (519/1254, 41.39%), followed by other effects (213/1254, 16.99%) and respiratory effects (182/1254, 14.51%). The other health categories had relatively lower proportions of tweets. For example, the cancer category (cancer is a health effect that is often associated with nicotine consumption) only took up 5.34% (67/1254) of the total tweets. Further hand-coding results showed that of the 200 randomly selected tweets, 40 (20%) mentioned that the health symptoms were a direct result of snus consumption or mentioned a negative opinion about snus. In addition, 28% (56/200) of the tweets discussed the harm reduction of snus, in terms of the health symptoms mentioned, when compared to smoking.

Discussion

Principal Findings

In our study, we showed that the proportion of snus-related tweets with a positive sentiment was significantly higher ($P<.001$) than the proportion of snus-related tweets with a negative sentiment. By using topic modeling, we observed that the positive sentiments toward snus might be the result of personal experiences and the perception that snus use is a safer alternative to smoking. In contrast, concerns about health risks might contribute to the negative sentiments in snus-related tweets. A further analysis showed that in snus-related tweets, the most popular health category was mouth effects, followed by other effects (eg, liver and skin effects) and respiratory effects.

Comparison With Previous Studies

Our temporal analysis showed an obvious peak in the number of snus-related tweets on October 3, 2021. After extracting all snus-related tweets from that day, we noticed that most of the tweets (67/100, 67%) discussed the possible use of snus by the son of a famous English former soccer player. This peak indicates the large impact of influencers on Twitter users.

Given that the top topic in snus-related tweets with a positive sentiment was related to switching from smoking to snus use, since snus was perceived as a safer option and there was no strong evidence in negative sentiment tweets indicating the gateway effect, it might be possible that Twitter users' perceptions on snus tend to lean toward the pathway hypothesis instead of the gateway hypothesis. This finding contradicts that of a focus group study, in which participants viewed snus use as a potential gateway to smoking [10]. There are 2 possible reasons for this inconsistency. First, the focus group was conducted in 2010, and the tweets used in our study were collected in 2021. It is possible that temporal differences might account for the difference in the perceptions of snus. Second, the conclusion from the focus group was based on a sample of 66 young adults who ranged in age from 18 to 26 years [10]. In comparison, our study included a broader range in terms of demographic characteristics, which may have led to the different results.

From the aspect of health risks, the health-related keywords identified in the tweets captured the majority of the potential health impact of snus. According to a report published by the Norwegian Institute of Public Health in 2019, the main potential adverse health effects of snus cover cancer, cardiovascular disease, mental disorders, and caries [25]. The health-related keyword frequency distribution from our study included these potential health effects through the oral, cardiovascular, cancer, and psychological effect categories, demonstrating the consistency between our findings from Twitter data and previous findings on the health risks of snus.

Limitations

Our study has several limitations. Data collected from Twitter may contain some bias. A study on tourist attraction visit sentiment data sourced from Twitter suggested that the tourists' sentiments could be affected by factors other than the tourist attraction itself, including the number of attraction sites that are visited in 1 day and whether the tourists are local visitors, out-of-state visitors, or international visitors [26]. Another study in 2012 suggested that the demographic distributions of Twitter users are different from those of the general population [27]. For example, around 31% of young adults who ranged in age from 18 to 24 years used Twitter, while this proportion was only 17% for adults aged between 25 and 34 years [27]. Therefore, our findings, which are based on Twitter data, may not represent the general population.

With regard to data collection and preprocessing, the keyword set we used may not have been comprehensive. For example, when collecting the data, we only included *snus* as the single keyword, which may have resulted in us missing some relevant tweets in our study. Additionally, in the processed data set, there might have still been some bot accounts, which can automatically deliver messages. This may have introduced some bias in our results. With regard to topic modeling, inferences based on keywords involve subjective judgments, even with the support of example tweets. In addition, the mentioning of health symptoms in snus-related tweets does not imply any causal relationship between snus and health risks. Our hand-coding results further validated this notion. Moreover, our study did not include the demographic information of Twitter users. Different demographic groups might perceive snus differently.

Conclusion

Our study showed more positive sentiments in snus-related tweets from Twitter users, which might have been due to the relative safety of snus when compared to that of smoking. Our study provided an efficient measurement of the public perceptions of snus among a relatively large sample by using social media data. According to the health belief model, the perceived susceptibility, seriousness, benefits, and barriers of actions explain health-related behaviors [28]. Therefore, these perceptions of snus are possibly a predictor of the public's snus consumption patterns. Our study will help policy makers better anticipate consumption behavior changes and make necessary policy changes. The results from our study will provide insights to policy makers on further regulations for snus. Future studies could take demographic and geographic factors into consideration to explore potential disparities in snus-related perceptions and discussions.

Acknowledgments

The research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health (NIH) and the Food and Drug Administration (FDA) Center for Tobacco Products under award number U54CA228110. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the FDA.

Data Availability

The data and scripts used for analysis and for creating the figures are available on request from the corresponding author (DL).

Authors' Contributions

ZX and DL conceived and designed this study. JC and SX analyzed the data. JC wrote the manuscript. ZX and DL assisted with the interpretation of analyses and edited the manuscript. All authors have approved the final article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Proportion of snus-related positive tweets over time.
[PNG File , 149 KB - medinform_v10i8e38174_app1.png]

Multimedia Appendix 2
Table S1. Topics mentioned in snus-related tweets with a neutral sentiment.
[DOCX File , 15 KB - medinform_v10i8e38174_app2.docx]

References

1. Smokeless tobacco product use in the United States. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/smokeless/use_us/index.htm [accessed 2022-06-07]
2. Clarke E, Thompson K, Weaver S, Thompson J, O'Connell G. Snus: a compelling harm reduction alternative to cigarettes. Harm Reduct J 2019 Nov 27;16(1):62 [FREE Full text] [doi: 10.1186/s12954-019-0335-1] [Medline: 31775744]
3. Directive 2014/40/EU of the European Parliament and of the Council of 3 April 2014 on the approximation of the laws, regulations and administrative provisions of the Member States concerning the manufacture, presentation and sale of tobacco and related products and repealing Directive 2001/37/EC. EUR-Lex. 2014 Apr 29. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32014L0040 [accessed 2022-08-04]
4. Tseveenjav B, Pesonen P, Virtanen JI. Use of snus, its association with smoking and alcohol consumption, and related attitudes among adolescents: the Finnish National School Health Promotion Study. Tob Induc Dis 2015 Oct 24;13:34 [FREE Full text] [doi: 10.1186/s12971-015-0058-3] [Medline: 26500472]
5. Foulds J, Ramstrom L, Burke M, Fagerström K. Effect of smokeless tobacco (snus) on smoking and public health in Sweden. Tob Control 2003 Dec;12(4):349-359 [FREE Full text] [doi: 10.1136/tc.12.4.349] [Medline: 14660766]
6. Bolinder G, Alfredsson L, Englund A, de Faire U. Smokeless tobacco use and increased cardiovascular mortality among Swedish construction workers. Am J Public Health 1994 Mar;84(3):399-404. [doi: 10.2105/ajph.84.3.399] [Medline: 8129055]
7. Lee PN. The effect on health of switching from cigarettes to snus - a review. Regul Toxicol Pharmacol 2013 Jun;66(1):1-5 [FREE Full text] [doi: 10.1016/j.yrtph.2013.02.010] [Medline: 23454227]
8. Timberlake DS, Huh J, Lakon CM. Use of propensity score matching in evaluating smokeless tobacco as a gateway to smoking. Nicotine Tob Res 2009 Apr;11(4):455-462. [doi: 10.1093/ntr/ntp008] [Medline: 19307445]
9. Ramström L, Borland R, Wikmans T. Patterns of smoking and snus use in Sweden: Implications for public health. Int J Environ Res Public Health 2016 Nov 09;13(11):1110 [FREE Full text] [doi: 10.3390/ijerph13111110] [Medline: 27834883]
10. Choi K, Fabian L, Mottey N, Corbett A, Forster J. Young adults' favorable perceptions of snus, dissolvable tobacco products, and electronic cigarettes: findings from a focus group study. Am J Public Health 2012 Nov;102(11):2088-2093 [FREE Full text] [doi: 10.2105/AJPH.2011.300525] [Medline: 22813086]
11. Smokeless tobacco products, including dip, snuff, snus, and chewing tobacco. U.S. Food & Drug Administration. URL: https://www.fda.gov/tobacco-products/products-ingredients-components/smokeless-tobacco-products-including-dip-snuff-snus-and-chewing-tobacco#stats [accessed 2020-04-17]
12. Myslín M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013 Aug 29;15(8):e174 [FREE Full text] [doi: 10.2196/jmir.2534] [Medline: 23989137]
13. Lu X, Chen L, Yuan J, Luo J, Luo J, Xie Z, et al. User perceptions of different electronic cigarette flavors on social media: Observational study. J Med Internet Res 2020 Jun 24;22(6):e17280 [FREE Full text] [doi: 10.2196/17280] [Medline: 32579123]
14. Bahreinifar S, Sheon NM, Ling PM. Is snus the same as dip? Smokers' perceptions of new smokeless tobacco advertising. Tob Control 2013 Mar;22(2):84-90 [FREE Full text] [doi: 10.1136/tobaccocontrol-2011-050022] [Medline: 21972063]
15. Van Rossum G. The Python Standard Library—Python 3.8.13 documentation. Python Software Foundation. 2020. URL: https://docs.python.org/3.8/library/ [accessed 2022-08-04]

16. Hutto CJ, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. 2014 Presented at: Eighth International AAAI Conference on Weblogs and Social Media; June 1-4, 2014; Ann Arbor, Michigan URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399
17. R: The R Project for Statistical Computing. R Foundation for Statistical Computing. URL: https://www.r-project.org/ [accessed 2022-08-04]
18. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. 2010 May Presented at: LREC 2010 Workshop on New Challenges for NLP Frameworks; May 22, 2010; Valletta, Malta p. 46-50 URL: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf [doi: 10.13140/2.1.2393.1847]
19. Bird S, Lope E, Klein E. Natural Language Processing with Python. Sebastopol, California: O'Reilly Media, Inc; 2009.
20. spaCy: Industrial-strength natural language processing in Python. Explosion. URL: https://spacy.io/ [accessed 2022-08-04]
21. Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. 2014 Jun Presented at: Workshop on Interactive Language Learning, Visualization, and Interfaces; June 27, 2014; Baltimore, Maryland p. 63-70 URL: https://aclanthology.org/W14-3110.pdf [doi: 10.3115/v1/w14-3110]
22. Chen L, Lu X, Yuan J, Luo J, Luo J, Xie Z, et al. A social media study on the associations of flavored electronic cigarettes with health symptoms: Observational study. J Med Internet Res 2020 Jun 22;22(6):e17496 [FREE Full text] [doi: 10.2196/17496] [Medline: 32568093]
23. Luo J, Chen L, Lu X, Yuan J, Xie Z, Li D. Analysis of potential associations of JUUL flavours with health symptoms based on user-generated data from Reddit. Tob Control 2021 Sep;30(5):534-541 [FREE Full text] [doi: 10.1136/tobaccocontrol-2019-055439] [Medline: 32709604]
24. Hua M, Alfi M, Talbot P. Health-related effects reported by electronic cigarette users in online forums. J Med Internet Res 2013 Apr 08;15(4):e59 [FREE Full text] [doi: 10.2196/jmir.2324] [Medline: 23567935]
25. Health risks from snus use. Norwegian Institute of Public Health. 2019. URL: https://www.fhi.no/en/publ/2019/health-risks-from-snus-use2/ [accessed 2022-07-09]
26. Padilla JJ, Kavak H, Lynch CJ, Gore RJ, Diallo SY. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. PLoS One 2018 Jun 14;13(6):e0198857 [FREE Full text] [doi: 10.1371/journal.pone.0198857] [Medline: 29902270]
27. Smith A, Brenner J. Twitter use 2012. Pew Research Institute. 2012 May 31. URL: https://www.pewresearch.org/internet/2012/05/31/twitter-use-2012/ [accessed 2022-06-06]
28. Rosenstock IM. Historical origins of the health belief model. Health Educ Monogr 1974 Dec 01;2(4):328-335. [doi: 10.1177/109019817400200403]

Abbreviations

CDC: Centers for Disease Control and Prevention
FDA: Food and Drug Administration
NIH: National Institutes of Health
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by T Hao; submitted 21.03.22; peer-reviewed by R Gore, J Li, A Dormanesh; comments to author 05.06.22; revised version received 20.07.22; accepted 22.07.22; published 29.08.22.

Please cite as:
Chen J, Xue S, Xie Z, Li D
Perceptions and Discussions of Snus on Twitter: Observational Study
JMIR Med Inform 2022;10(8):e38174
URL: https://medinform.jmir.org/2022/8/e38174
doi:10.2196/38174
PMID:36036970

Original Paper

Search Term Identification Methods for Computational Health Communication: Word Embedding and Network Approach for Health Content on YouTube

Chau Tong[1], PhD; Drew Margolin[1], PhD; Rumi Chunara[2,3], PhD; Jeff Niederdeppe[1,4], PhD; Teairah Taylor[1]; Natalie Dunbar[5]; Andy J King[6,7], PhD

[1]Department of Communication, Cornell University, Ithaca, NY, United States
[2]Department of Biostatistics, School of Global Public Health, New York University, New York, NY, United States
[3]Department of Computer Science & Engineering, Tandon School of Engineering, New York University, New York, NY, United States
[4]Jeb E Brooks School of Public Policy, Cornell University, Ithaca, NY, United States
[5]Greenlee School of Journalism and Communication, Iowa State University, Ames, IA, United States
[6]Cancer Control and Population Sciences, Huntsman Cancer Institute, Salt Lake City, UT, United States
[7]Department of Communication, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Chau Tong, PhD
Department of Communication
Cornell University
494 Mann Library
Ithaca, NY, 14850
United States
Phone: 1 608 334 9909
Email: ctt39@cornell.edu

Abstract

Background: Common methods for extracting content in health communication research typically involve using a set of well-established queries, often names of medical procedures or diseases, that are often technical or rarely used in the public discussion of health topics. Although these methods produce high recall (ie, retrieve highly relevant content), they tend to overlook health messages that feature colloquial language and layperson vocabularies on social media. Given how such messages could contain misinformation or obscure content that circumvents official medical concepts, correctly identifying (and analyzing) them is crucial to the study of user-generated health content on social media platforms.

Objective: Health communication scholars would benefit from a retrieval process that goes beyond the use of standard terminologies as search queries. Motivated by this, this study aims to put forward a search term identification method to improve the retrieval of user-generated health content on social media. We focused on cancer screening tests as a subject and YouTube as a platform case study.

Methods: We retrieved YouTube videos using cancer screening procedures (colonoscopy, fecal occult blood test, mammogram, and pap test) as seed queries. We then trained word embedding models using text features from these videos to identify the nearest neighbor terms that are semantically similar to cancer screening tests in colloquial language. Retrieving more YouTube videos from the top neighbor terms, we coded a sample of 150 random videos from each term for relevance. We then used text mining to examine the new content retrieved from these videos and network analysis to inspect the relations between the newly retrieved videos and videos from the seed queries.

Results: The top terms with semantic similarities to cancer screening tests were identified via word embedding models. Text mining analysis showed that the 5 nearest neighbor terms retrieved content that was novel and contextually diverse, beyond the content retrieved from cancer screening concepts alone. Results from network analysis showed that the newly retrieved videos had at least one total degree of connection (sum of indegree and outdegree) with seed videos according to YouTube relatedness measures.

Conclusions: We demonstrated a retrieval technique to improve recall and minimize precision loss, which can be extended to various health topics on YouTube, a popular video-sharing social media platform. We discussed how health communication

scholars can apply the technique to inspect the performance of the retrieval strategy before investing human coding resources and outlined suggestions on how such a technique can be extended to other health contexts.

(JMIR Med Inform 2022;10(8):e37862) doi:10.2196/37862

KEYWORDS

health information retrieval; search term identification; social media; health communication; public health; computational textual analysis; natural language processing; NLP; word2vec; word embeddings; network analysis

Introduction

Background

Researchers are increasingly interested in understanding the types and accuracy of health-related messages produced in the public communication environment (PCE) [1-5]. Given the proliferation of web-based health information sources and social media platforms in which people generate, share, and access information [6], identifying and capturing what message content individuals are likely to see when looking for information about health (ie, seeking), as well as what information people might encounter while being on the web (ie, scanning) [7-9], is crucial in gaining insights into issues, including misinformation or inequities, on web-based platforms within the larger PCE.

Nevertheless, identifying appropriate strategies to retrieve this information is challenging. To gather data for analysis, researchers often rely on the standard approach of searching for content using keywords, which usually involve a set of technical (eg, medical) terms that describe a condition or behavior of interest (eg, “colon cancer” or “diabetes”) [10-12]. However, keyword search strategies that are solely based on technical concepts cannot account for the multifaceted nature of web-based information. A primary reason is that the messages in the contemporary PCE are often generated by users and, thus, often include colloquial terminology rather than medical terminology [7,13-15]. This phenomenon has been well documented in consumer health vocabularies research, which examines the language gap between official medical texts and user-generated content, such as question and answer (Q&A) sites (Yahoo! Answers) and social media platforms (eg, Twitter) [16-19].

In addition to messages that do not include technical keywords, another type of content that might be overlooked by the standard retrieval approach is what could be categorized as content that misleads by omission (eg, messages that describe risky behaviors but fail to name the medical risk it exposes an individual to) [20-22]. For example, messages promoting a fad diet, which might be associated with a specific medical condition but do not mention this risk nor the condition itself, will not be retrieved by keywords naming the condition.

Failure to retrieve these messages could result in the biased identification of content, especially in light of research showing how search results vary according to specific queries [23] and how social media language varies across different geographical locations [24]. In other words, retrieving (and analyzing) only messages produced with the “official” technical language can lead researchers to overlook the information consumed and barriers faced by underprivileged groups [25,26] or users who lack the skills and knowledge to correctly use official medical vocabularies to access information [27,28]. For these reasons, public health researchers trying to understand the PCE would benefit from a principled, replicable process for searching for web-based content relevant to medical terms but not exclusively restricted to them. Such a process would also inform web-based users’ health information–seeking efforts by enabling the retrieval of health-related information from commonly used slang or nontechnical queries.

This paper proposes such a retrieval process for YouTube. Using the platform’s application programming interface (API) to retrieve videos and the inferred relatedness between videos determined by YouTube’s proprietary algorithm, our process retrieves videos that (1) are frequently relevant to understanding the PCE related to a focal technical term, (2) are distinct from the videos retrieved directly with the focal term, and (3) can be easily distinguished from irrelevant videos that could otherwise absorb researchers’ attention. Such a search identification approach balances the trade-off between recall and precision [29], identifying content that would not have been found using typical keywords without requiring human coders to sift through large quantities of irrelevant content.

In the following sections, we summarize relevant research on PCE content retrieval, highlighting strengths and weaknesses. We then discuss the rationale for using YouTube before detailing the techniques used to identify relevant content beyond formal medical concepts. We illustrate the techniques using cancer screening as a case study. We conclude with a discussion of the potential for application of the technique across other topics and platforms.

Challenges of Health-Related Vocabulary Inconsistencies

User-generated health content presents important challenges to researchers attempting to retrieve content from this environment, particularly as (1) researchers may not know the vocabulary users use to discuss health topics and (2) users can mislead each other by failing to mention relevant information.

Research has shown that patients often do not conceptualize diseases, treatments, or risks in the same terms as health care practitioners [30-32]. Most plainly, the literature on consumer health vocabulary [15-17] shows that the terms used by laypeople are different from those used by health care practitioners. For example, questions about health topics posted on Q&A sites (eg, Yahoo! Answers and WebMD) by laypeople were found to contain misspelled words, descriptions, and background information and were more colloquial than texts by health professionals [13,33]. A more recent example is the COVID-19 pandemic, where infodemiologists identified a

variety of terms using Google Trends that referred to the virus, including "stigmatizing and generic terms" (eg, "Chinese coronavirus" and "Wuhan virus") that had not been identified by other research using more agreed upon and technical language about the virus [34]. These works suggest that user vocabulary, which is distinct from medical vocabulary, is important for understanding how individuals conceive of their health and the medical vocabulary related to it when looking for or coming upon health information on the web. More broadly, these different terminologies can reflect different ways of conceptualizing health issues [32,35,36].

It is not surprising then that user vocabulary is important for identifying relevant health-related posts on social media, as research indicates that retrieval performance significantly changes when users' health queries are reformulated using formal, professional terminologies [23]. Thus, if researchers do not know what the user vocabulary is for a given topic, their retrieval strategy will be biased to identify only content posted by users who use technical medical vocabulary. Moreover, this bias is unlikely to be neutral with respect to larger public health concerns. In particular, differences of this nature, such as conceptualization of illness and preferred vocabulary, have been shown to be associated with important differences in outcomes [25,26,37]. Such conceptual differences would likely manifest in differences in user vocabulary.

Problems of Omission in Health Information Retrieval

Another weakness of retrieving user-generated health messages with technical terms is that this strategy cannot, by definition, identify information that omits that term. However, this failure to connect risks to outcomes can be precisely what makes user-generated content misleading. It is well established that many people lack broad knowledge about risk factors for many leading causes of death in the United States and beyond [38-40], and people routinely receive information that fails to link common risk factors and behaviors to negative health outcomes [41]. Perhaps the best known (and most damaging) example is the failure of tobacco companies to mention that cigarette smoking causes cancer in their promotional materials [42]. This misrepresentation by omitting and distancing from medical terms (eg, disease) is common for unhealthy products (eg, alcohol) [43].

In such cases, the PCE misleads by omission as it fails to assign the appropriate words to what is medically accurate in the offline world. This has the potential to mislead the public and makes relevant messages hard to find, as their relevance (to researchers) is defined by what is absent (the mention of the risk). An example is the "Tide Pod challenge" that emerged in 2017 as a popular internet trend. The Tide Pod challenge is dangerous as it fails to connect the terms "Tide detergent" and "eat" with the concept (or concept family) of "poison." A trained medical professional would not discuss "eating" Tide Pods without also mentioning the danger, although users can (and did) do so. Such misleading (and dangerous) user messages cannot be retrieved by strategies that focus on the harm—poisoning.

In the case of well-researched and widely understood risks, such as the connection between cigarette smoking and lung cancer, this weakness can be overcome by simply naming the risk factor (ie, searching for "lung cancer"). However, to restrict searches to known and well-documented high-risk behaviors is to again return researchers to their cultural bubble [44]. As evidenced by the emergence of the Tide Pod challenge, user-generated content can be extraordinarily inventive, creating new risky behaviors unknown to the medical community. For example, dangerous fad diets cannot be identified by searching for the risks they pose. Instead, what is needed is a way of identifying vocabulary that is "near" to the condition of interest, broadening the net so that researchers can identify messages misleading by omission.

For both reasons, researchers should find ways to escape the strictures of official, technical vocabulary when retrieving information to characterize the PCE. Researchers instead need search terms that include culturally relevant colloquial terms that are related to medical terms and terms that identify behaviors or practices *in the neighborhood* of medical terms but which can identify content when those terms are omitted.

YouTube as Public Health Information Source and Site of Inquiry

In this study, we focus on YouTube videos as a meaningful message source of the PCE. We selected YouTube for 2 reasons. First, YouTube is one of the most widely used web-based social media and content platforms [45]. Second, YouTube has become increasingly relevant as a source of health information. With its dual function as a reservoir of video content and a social networking platform in which users acquire information through interactions with the content and fellow users, YouTube has served as an informational resource for learning about diverse health topics for users [46,47].

Extant research on medical and health information on YouTube suggests several issues with the quality of YouTube content. A meta-analysis found that YouTube videos tend to prevalently contain misinformation, an implication of which is the potential of the platform to alter beliefs about health interventions [46]. A limitation of these studies (and a weakness shared by many YouTube studies) is the search strategies used to identify relevant content. To address this gap in current research, our project aims to answer 2 research questions (RQs).

The first main RQ asks the following: for a given medical or health term of interest (ie, a focal term for retrieval), does our proposed search term identification strategy retrieve health messages that are relevant to understanding the public health communication environment related to that seed term and do not explicitly use that term (such that the traditional medical or technical search terms would have failed to retrieve them)? To provide a satisfactory answer to this question, a search strategy must (1) retrieve content relevant to the seed term (called precision) and (2) find relevant content that is novel, (ie, different from what would be returned by the seed term alone, called recall), without sacrificing too much precision. This leads to our second RQ: can the derived strategy identify relevant, novel messages with sufficient precision to be practically useful?

Methods

Rationale for Cancer Screening Focal Terms

Cancer is one of the biggest public health issues in the United States and, thus, is a topic that requires meticulous attention from multiple stakeholders, including public health practitioners and communicators. A particular challenge to the prevention and management of various cancer types is the persistent disparities in screening, incidence, and mortality rates across different population groups [48]. Given the significance of cancer and the important implications of cancer screening disparities, we chose cancer screening as the subject of examination in this paper.

To this end, we first demonstrate our methodological technique using the primary colorectal cancer screening option—"colonoscopy"—as our focal term. Colorectal cancer is the third most diagnosed and third most deadly cancer in the United States, which disproportionately affects Black individuals compared with non-Hispanic White Americans [49]. We then replicate the analyses using other cancer screening tests (fecal occult blood test, mammogram, and pap test) as focal terms to illustrate how the technique performs in other cancer contexts, including breast and cervical cancer.

Retrieving YouTube Videos From the Focal Term

We collected data from YouTube via the YouTube API (version 3). Using the "search: list" end point (used for the search function) allowed us to retrieve 2 types of data: videos that are most relevant to a search query or set of queries (the "q" parameter with "relevance" sorting) and videos that are related to a specific or set of videos (the "related-to-video-id" parameter) according to YouTube algorithms [50]. We note that collecting data through this API approach bypasses localization and personalization—factors that play important roles in search results that are presented to specific individuals. As our purpose is to demonstrate a methodology that can be systematically extended to other contexts in future research, we deem this approach to be appropriate in giving us the results as close to a default setting as possible.

On August 22, 2021, using the YouTube Data Tools software [51], we retrieved a set of 250 videos most relevant to the search term "colonoscopy." These 250 videos comprise our core set. In addition, we retrieved 4304 videos "related to" this core set, which gave us 4554 videos in total in the initialization set. We retrieved these videos' unique identifiers, text data (video titles and descriptions), and metadata (publication date and engagement statistics).

Word Embeddings

Word embedding is an unsupervised method of learning word vectors using a neural network model [52]. The basic aim of word embeddings is to identify words that appear in "similar contexts" as the focal term. The technique calculates a proximity score; that is, the extent to which 2 terms are near to one another in a multidimensional space. This score acts as a measure of "semantic similarity." Thus, it is a useful way of finding texts that discuss a particular concept without explicitly mentioning it. Texts that mention a word's close neighbors (in the multidimensional space) are likely talking about ideas where that word is relevant as well, even if the word itself is not there. We used word embeddings to find YouTube content that is relevant to "colonoscopy" but which may not mention the word itself.

We applied word embeddings using the word2vec approach to the text data of our initialization set of 4554 videos. Specifically, we used the text of the 4554 video titles and descriptions to build a corpus. Subsequently, after preprocessing and standardization steps (including removal of emojis, signs, and stop words; performing lowercasing; converting text to American Standard Code for Information Interchange; encoding; and removing leading or trailing spaces), a word2vec model was trained on the text to identify the terms with the most semantic similarity to the term "colonoscopy" (word2vec R package) [53].

We then used the top 6 "nearest neighbors" to "colonoscopy" as new search terms to retrieve more videos (250 videos for each neighbor) to inspect the new content.

Human Coding and Natural Language Processing to Evaluate Recall Improvement

The goal of retrieving new content from the nearest neighbors is the improvement of recall over a direct search—the identification of videos that are relevant to "colonoscopy" but which would not be found by searching directly for it. To assess this recall improvement, we took a random 10% (150/1500) sample (25 videos for each neighbor) and coded them for relevance. Coding was done by a research team member (AJK, the paper's last author) with expertise in cancer control and cancer communication.

Specifically, a video was coded as relevant if the video content contained (1) any aspect of screening preparation or procedures (eg, bowel preparation, personal experiences, and clinical discussions) or (2) general information on colorectal cancer or colorectal cancer screening in terms of cancer prevention or early detection. This included content where a patient underwent a colonoscopy but perhaps for a chronic condition (eg, ulcerative colitis or Crohn disease). Obscure terms identified through this process were also looked up as needed to confirm relevance (eg, "suprep"—a commercial brand for a bowel preparation kit).

We evaluated recall in 2 ways. First, we assessed how many of the relevant "found" videos would have been identified using the search term alone. We did this by counting the number of relevant videos in the newly found set containing the term "colonoscopy." Those that did not contain "colonoscopy" but were nonetheless relevant to it constituted a recall improvement. Second, we examined whether these newly found videos were substantively different—in terms of contents, topics, and focus—from the core set. Using the R package *quanteda* [54], we calculated the average Euclidean distances between the text features embedded in the different video sets. Euclidean distance is a pairwise distance metric that measures dissimilarities between the text features in different corpora. We then used hierarchical clustering analysis, with the complete linkage method (*hclust* function in *stats* version 3.6.2), to determine

whether videos in different sets were substantially overlapping in content.

Network Analysis to Evaluate Precision

Strategies to improve recall are often offset by a substantial loss of precision. In our case, although the nearest neighbors may retrieve many more relevant videos, they could, at the same time, bring in many irrelevant videos. This introduces the risk of increasing human coding costs or other resource-intensive techniques of classification. Such precision loss needs to be mitigated so that it occurs at a manageable level. To implement this, we used the "related to video id" API end point, which reports whether a set of videos are "related to" the others (zero crawl depth), to query the relationships between the new videos retrieved from the top neighbor terms and the colonoscopy videos from the core set. Specifically, if video A is related to video B in a set, there is a connection (or link) between them. These relations were used to create a network with videos being nodes and the connections between them being edges.

We then calculated 3 network measures of relatedness: indegree (videos from the core set linking to a newly found video), outdegree (videos in the core set linking to each newly found video), and total degree (sum of indegree and outdegree). We expected that the newly found irrelevant videos would have few, if any, links to the videos known to be about "colonoscopy," whereas videos with even loose relevance would have at least some connections to the core set. To examine the extent to which these degree scores were associated with relevance (according to human coding), the corresponding precision and recall statistics at different degree levels were inspected. If our technique worked effectively, there would be some threshold of degree—the number of connections between a newly found video and the core set—at which videos with this degree or higher are not only reasonably novel (improving recall over the core set) but also reasonably relevant (maintaining precision at a manageable level).

Ethics Approval

This study did not involve the use of human subjects, as the data collected were strictly limited to publicly available data on YouTube; therefore, no ethics approval was applied for. This rationale is consistent with the institutional policies where the research was conducted.

Results

Word Embeddings

Table 1 provides the list of neighbor terms to the focal term "colonoscopy" and their ranks based on semantic similarity, according to word embedding results.

A visual inspection suggests these nearest neighbor terms fit our goals for this method: they contain nontechnical terms (eg, "cleanse" or brand names such as "plenvu") that are relevant to colorectal health. We selected the top 6 terms ("suprep" to "miralax"), retrieved an additional 1500 videos (250 each), and coded a subset of 10% (150/1500 random videos) for the recall analysis.

Table 1. Neighbor terms to "colonoscopy" and similarity scores.

Term[a]	Similarity score	Rank
"suprep"	0.9722890	1
"peg"	0.9519246	2
"sutab"	0.9513488	3
"plenvu"	0.9504289	4
"glycol"	0.9498276	5
"miralax"	0.9449067	6
"rectal"	0.9435940	7
"cleanse"	0.9422708	8
"cologuard"	0.9421358	9
"colorectal"	0.9403084	10

[a]Neighbor terms are terms with the most semantic similarity (with corresponding high similarity scores or low ranks) to "colonoscopy" based on YouTube video data. Score refers to the cosine similarity metric between word embeddings (ie, terms) in a multidimensional vector space.

Human Coding and Natural Language Processing to Evaluate Recall Improvement

Table 2 displays the retrieval statistics, of which 34% (51/150) of the coded videos were deemed relevant. More importantly, of these 51 videos, 21 (41%; 21/150, 14% of the coded sample) did not contain the term "colonoscopy," meaning that identifying them improved recall over what would have been found simply by searching for "colonoscopy." This supported our expectation that the word embedding approach helped address the recall problem inherent in using technical language.

We next assessed whether these newly found videos were substantively different—in terms of contents, topics, and focus—from what would be retrieved with the typical strategy. To assess this, we compared the Euclidean distances between textual features of the core set (250 videos) with those of the newly found videos (Table 3). Here, higher values meant greater distance. For example, the distance between "miralax" and

"peg" was the smallest among our groupings, indicating that videos in these 2 sets shared the most similar words compared with other pairs.

Table 2. Retrieval statistics in the sampled videos for the top 6 neighbors of "colonoscopy."

Terms	Sample of coded videos, N	Relevant (precision), n (%)	Relevant and mention of "colonoscopy," n (%)	Relevant and does not mention "colonoscopy" (recall improvement), n (%)
"suprep"	25	18 (72)	9 (36)	9 (36)
"peg"	25	1 (4)	0 (0)	1 (4)
"sutab"	25	4 (16)	4 (16)	0 (0)
"plenvu"	25	23 (92)	15 (60)	8 (32)
"glycol"	25	0 (0)	0 (0)	0 (0)
"miralax"	25	5 (20)	2 (8)	3 (12)
Total	150	51 (34)	30 (20)	21 (14)

Table 3. Euclidean distance between the text features of original "colonoscopy" video set and video sets generated from top 6 neighbor terms[a].

Term	1	2	3	4	5	6
"colonoscopy"	0	255.61	257.97	241.5	248.9	254.68
"miralax"	N/A[b]	0	6.32	20.1	21.8	7.14
"peg"	N/A	N/A	0	22.2	23.1	6.86
"plenvu"	N/A	N/A	N/A	0	20.6	19.08
"suprep"	N/A	N/A	N/A	N/A	0	20.57
"sutab"	N/A	N/A	N/A	N/A	N/A	0

[a]Cell values indicate dissimilarities of the text features belonging to any pair of video sets. Larger values indicate larger distances, and 0 indicates identical text features. "Glycol" was removed because of 0 relevant videos retrieved.

[b]N/A: not applicable.

Relative frequency analysis was used to further illustrate these differences by highlighting the differences in the text features of the core set as opposed to the newly found set. As Figure 1 shows, words such as "colonoscopy," "dr," "preparing," "colon," and "polyp" were disproportionately more likely to occur in the core set, whereas words such as "suprep," "prep," "kit," "bowel," and "miralax" were distinct terms found in the newly found set.

Hierarchical agglomerative clustering performed on the text features of the newly found set and the core set (using the complete link method) revealed that the text features in the videos retrieved from neighbor terms (newly found set) were more similar to such from other neighbor terms than to the core set (Figure 2). In other words, these results show that our approach helped identify videos that are relevant to "colonoscopy" without including the term itself (ie, improving recall); furthermore, these newly found relevant videos additionally enhanced the topical diversity of our retrieved data (by focusing on preparation brands and procedures).

Figure 1. Relative frequencies of words in the colonoscopy video set and the combined top 5 neighbor term video set. Words that are "key" to each video set were plotted. Original: the set of videos found with the search query "colonoscopy." Reference: the set of videos found with 5 nearest terms to "colonoscopy" ("suprep," "peg," "sutab," "plenvu," and "miralax"). chi2: chi-square value.

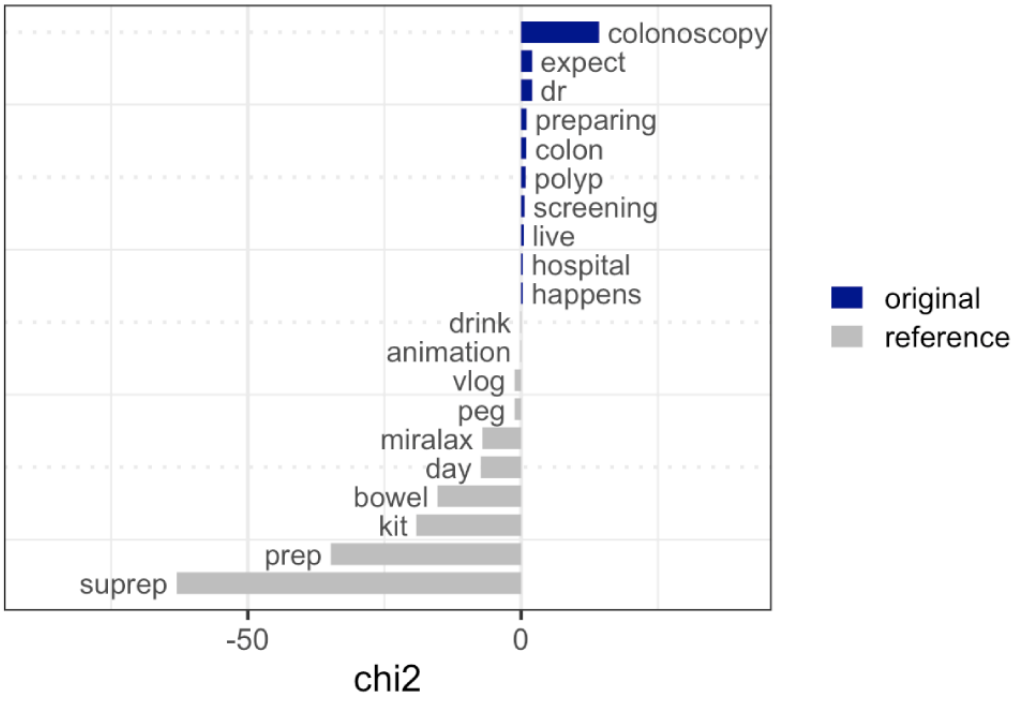
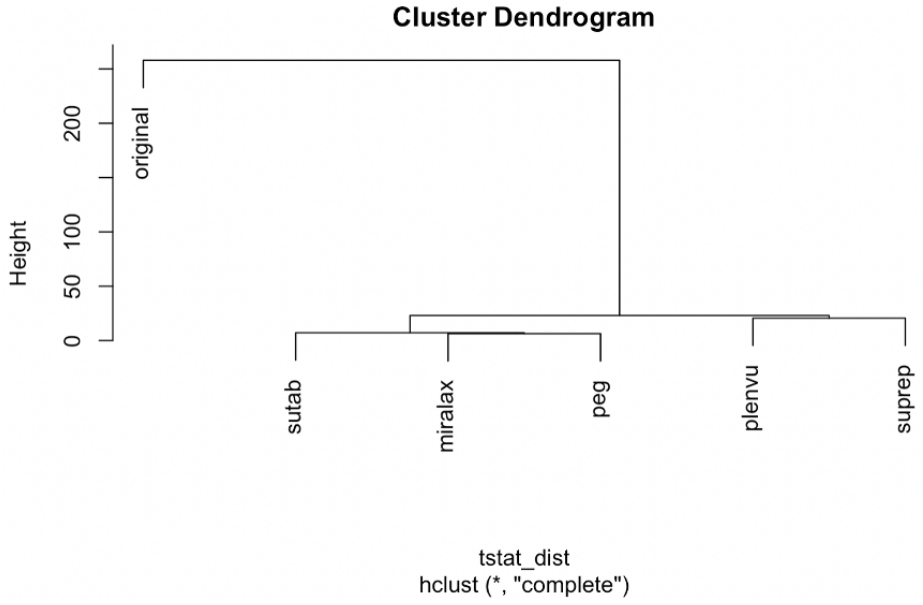


Figure 2. Visualization of distances between video sets. Hierarchical cluster analysis indicating dissimilarities and distances between original (set of videos found with the search query "colonoscopy") and sets of videos found with 5 nearest terms to "colonoscopy" ("suprep," "peg," "sutab," "plenvu," and "miralax").



Network Analysis to Evaluate Precision

Table 4 shows the results of the comparison between a found video's degree of connection to the core set and its associated relevance according to human coding. We first note that new videos that are in other languages than English (28/150, 18.7%) were found to have no connections with the core set videos. To avoid having this add bias to our results, we excluded these 28 videos, as well as 8 videos that were already found in the original set and 1 video where YouTube returned missing metadata (37/150, 24.7% excluded in total). We then performed a comparison on the remaining 75.3% (113/150) of videos (the final total in the "cumulative count of videos," also the denominator).

Table 4. Relevance of newly found videos by the number of links to the original set of colonoscopy videos (total degree).

Total degree[a]	Count of videos with total degree, N	Number of videos coded as "relevant" (relevancy), n (%)	Cumulative count of nonduplicate videos, N	Cumulative count of nonduplicate relevant videos, n	Cumulative precision[b] (%)	Cumulative recall[c], (%)	Cumulative F_1-score[d], (%)
44	1	1 (100)	1	1	100	2.7	5.3
41	1	1 (100)	2	2	100	5.4	10.3
26	1	1 (100)	3	3	100	8.1	15.0
23	1	1 (100)	4	4	100	10.8	19.5
22	1	1 (100)	5	5	100	13.5	23.8
21	1	1 (100)	6	6	100	16.2	27.9
20	2	2 (100)	8	8	100	21.6	35.6
19	1	1 (100)	9	9	100	24.3	39.1
18	1	1 (100)	10	10	100	27.0	42.6
17	2	2 (100)	12	12	100	32.4	49.0
16	1	1 (100)	13	13	100	35.1	52.0
15	2	2 (100)	15	15	100	40.5	57.7
14	1	1 (100)	16	16	100	43.2	60.4
13	1	1 (100)	17	17	100	45.9	63.0
12	2	2 (100)	19	19	100	51.4	67.9
11	2	2 (100)	21	21	100	56.8	72.4
10	1	1 (100)	22	22	100	59.5	74.6
9	1	1 (100)	23	23	100	62.2	76.7
7	2	1 (50)	25	24	96	64.9	77.4
6	1	1 (100)	26	25	96	67.6	79.4
5	2	0 (0)	28	25	89	67.6	76.9
4	2	2 (100)	30	27	90	73.0	80.6
3	2	1 (50)	32	28	88	75.7	81.2
2	5	1 (20)	37	29	78	78.4	78.4
1	5	1 (20)	42	30	71	81.1	75.9
0	71	7 (10)	113	37	33	100	49.3

[a]The sum of connections each new video has with the videos in the original colonoscopy video set.

[b]The cumulative count of relevant videos divided by the cumulative count of all videos.

[c]Cumulative count of relevant videos divided by the total number of new and nonduplicate 37 relevant videos.

[d]The harmonic mean of cumulative precision and cumulative recall.

The first 4 columns in Table 4 show the total degree (number of connections) and counts of videos with corresponding total degrees in comparison with the relevance statistics. Specifically, all videos with a total degree >7 had been coded as relevant, meaning precision is 100% at or above this threshold. More importantly, although precision was imperfect below this threshold, it remained very high. In fact, when we examined videos of degree ≥1, we found that 71% (30/42) had been coded as relevant. This means that a human coding team choosing to use this liberal threshold (at least one connection to any video in the core set) for choosing videos to code would see >2 relevant videos for every irrelevant one, thus expending limited resources examining irrelevant videos.

The cumulative columns on the right-hand side of the table display the trade-offs that would face a coding team. The cumulative count of relevant videos adds up to 37, which is the 51 coded as relevant (Table 2) excluding the 8 videos already found in the original data set (as reported above) and 6 non-English videos that had been coded as relevant. Cumulative precision refers to the relevance of the videos at or above this threshold. Cumulative recall shows the portion of the relevant videos in the set that are preserved at this threshold. As the threshold tightens, precision improves (irrelevant videos are discarded) but recall declines (some relevant videos are discarded too). For example, if a team chose to examine videos with at least three connections to the core set (degree ≥3), they would find 32 videos, 28 of which are relevant (88% precision),

and miss out on only 9 of the 37 possible (75.7% recall). In other words, this technique provides a basis for researchers to inspect the performance of the retrieval strategy before investing human evaluation and coding resources.

Replication: Other Cancer Screening Tests

We extended our analyses to 3 additional focal terms to illustrate the breadth of the technique's applicability. The first, "FOBT," refers to the fecal occult blood test, another screening method for colorectal cancer. The second and third are "mammogram" and "pap test," screening tests for breast cancer and cervical cancer, respectively. We chose cancer screening as an illustrative case as these are common cancer types that are often discussed on social media [3,55] such that research would benefit from identifying relevant content that does not explicitly mention these technical, formal screening tests.

As shown in the summary statistics in Table 5, the results for these terms were comparable with "colonoscopy." For each focal term, searches using the nearest neighbor terms uncovered through word2vec identified a wide range of new videos that were distinct from the original sets, improving recall (see Multimedia Appendix 1 for dissimilarity measures of new vs original content). Similar to the results for "colonoscopy," filtering videos based on their degrees of connection to the core set (for the respective focal term) improved precision while maintaining reasonable recall. For both "FOBT" and "pap test," researchers could inspect only videos with a degree of ≥1 and would find a few irrelevant videos while maintaining most of the new videos in the set. For "mammogram," the recall statistics of videos with at least one connection is lower (30%); however, even if researchers chose to drop this filter and inspect all videos, they would find that approximately 1 in 3 new videos found is relevant. Thus, researchers would not be at risk of being overwhelmed with irrelevant content.

Table 5. Summary retrieval statistics for "colonoscopy," "FOBT," "mammogram," and "pap test."

Focal term	Top nearest neighbor terms	Sample of coded videos (videos per term)	New and nonduplicate relevant videos (set A), N	Videos with degree ≥1 (set B)[a], N	Videos with degree ≥1[a] and coded as new and relevant, n (A∩B)	Precision, n/N (%)	Recall, n/N (%)
Colonoscopy	• "suprep" • "peg" • "sutab" • "plenvu" • "glycol" • "miralax"	150 (25)	37	42	30	30/42 (75)	30/37 (81)
FOBT[b]	• "iFOBT" • "hemosure" • "immunochemical" • "immunostics" • "guaiac"	125 (25)	50	33	27	27/33 (82)	27/50 (54)
Mammogram	• "smartcurve" • "breastcheck" • "biopsy" • "ultrasound" • "breastcancerawareness"	250 (50)	77	28	23	23/28 (82)	23/77 (30)
Pap test	• "Colposcopy" • "Smear" • "ASCUS"[c] • "papsmear" • "STD"[d]	250 (50)	87	65	59	59/65 (91)	59/87 (68)

[a]Videos with at least one connection to the original set of videos resulted from the focal terms.

[b]FOBT: fecal occult blood test.

[c]ASCUS: atypical squamous cells of undetermined significance.

[d]STD: sexually transmitted disease.

Discussion

Principal Findings

This paper proposes a novel approach to improving the retrieval of user-generated health content. Using medical concepts as focal terms, we used the similarity-based word embedding approach to detect new search terms related to focal terms but not restricted to technical vocabulary. In line with previous research using similar methods (eg, word, sentence, or biomedical term embeddings), we identified less widely known terms in user-generated public discourse related to cancer screening tests. Quantitative textual analysis of the newly discovered content returned from the top neighbor terms

indicated that these videos were distinct from the original video sets in terms of lexical and topical foci. Network analysis showed that retrieval precision can be improved by detecting videos with at least one total degree; that is, those with at least one connection to others in the same networks. Researchers could use the technique to inspect the performance of their retrieval strategy before investing additional evaluation resources [56,57]. Beyond suggesting the value of this technique, our analyses provide insights into specific message gaps if user-generated vocabulary is overlooked.

First, our results indicate that commercial speech, particularly tagged by brand names such as "suprep" and "miralax," was particularly prominent and useful for identifying relevant content. In essence, users produced and consumed videos about "prepping," which could be used for colonoscopies, in reference to branded products. This raises an important follow-up question—do these videos provide accurate information? As reviewed previously, the history of corporate actors misleading consumers by omission of risks is substantial [58,59]. Although this would be an analysis for further study, we point out here the importance of retrieving information about medical topics using commercial terms rather than just medical or technical terms.

Second, we note that our results did not provide examples of de novo slang synonyms (akin to "the sugars"). Rather, when users created terms, they were more likely to be portmanteaus of simple vocabularies, such as "breastcheck," "papsmear," or even "breastcancerawareness." This merging of words into one term is unsurprising insofar as it is consistent with the conventions for the creation of hashtags; however, this should serve as a caution to researchers to consider these nonstandard constructions in their retrieval strategies. In other words, for the terms searched in this study, we found little evidence of colloquial language. However, for any health topic, there is the possibility that such language is used in less intuitive ways. Although we did not find that to be the case for our focal terms, the possibility exists, and this technique could have the potential to identify such in other cases.

More broadly, our analysis reveals that although user-generated vocabulary can often be sensibly interpreted after the fact (Plenvu's website advertises it as a colonoscopy prep technique, and "breastcheck" is intuitively related to breast cancer), the most common terms are not always easy to guess in advance, that is, before analyzing some data. This observation supports the arguments that motivated this research, suggesting that researchers should first learn how users talk about medical topics and then create retrieval strategies to build fuller data sets for analysis of what they are saying. Although we do not have explicit evidence here that vocabularies are associated with particular social groups, or, in particular, marginalized groups, the presence of corporate brand names suggests, at the very least, that targeted marketing efforts could play such a role for particular medical topics. This is a topic for further research.

Limitations

There are several limitations to this study. First, our analysis focused only on cancer screening tests as focal terms because of this project's inclusion in a larger project focusing on colorectal cancer screening information in the PCE. Our purpose was to demonstrate a methodological technique in the context of cancer with the understanding that future research will need to assess any unique challenges that might apply to noncancer screening health topics or medical terminologies of interest (eg, vaccines or information about diabetes management). Although we see no methodological reasons why this technique could not be applied to other keywords and terminologies, future research would be needed to support this expectation.

The second limitation is that the word embedding model was trained on YouTube textual content, and our technique relied on YouTube's relatedness data to distinguish between relevant and irrelevant videos. This means that the effectiveness of the present approach is limited to YouTube. Although there are good reasons to start with YouTube as a prevalent source of health-related information, we encourage future research to consider developing similar approaches for other domains where user-generated texts are found on the web, including websites, Q&A forum posts, and other social networking sites [21,57]. Importantly, many specific techniques may not be exportable from platform to platform. For example, although YouTube tracks relatedness between videos, messages on Twitter are often related by hashtags. Thus, rather than searching for relevant neighbor words, researchers might focus on identifying relevant neighbor hashtags. In Q&A forums or other content with threaded replies, researchers might incorporate this hierarchical information to identify the most relevant content (eg, terms used in top-level posts).

A final limitation is that conducting this process requires some familiarity with available natural language processing and computational tools. We believe the increasing application of computational methods in social science research, as well as the proliferation of training in R and Python languages for social scientists, increases the likelihood that this technique could be used by those with limited natural language processing proficiency. Nevertheless, health communication is an inherently interdisciplinary field in which we see great potential for collaborations among communication scientists, public health and medical researchers, and data scientists. However, future work might strive to make this technique more accessible through the creation of specific tools and materials to assist health communicators and public health professionals in applying these approaches in future health promotion and education efforts.

Conclusions

This study demonstrated the potential of using similarity-based word embedding techniques for computational health communication research to improve recall and maintain precision in retrieving content that could be overlooked by standard medical terminologies. The study reveals that there are indeed relevant messages to medical topics in the PCE that do not use medical vocabulary, and that many of these can be identified. Although the impact of overlooking these messages on health disparities cannot be determined, these results suggest that further study in this area is warranted.

Acknowledgments

This work was supported by the National Cancer Institute of the National Institutes of Health under award number R37CA259156. 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

None declared.

Multimedia Appendix 1
Supplementary replication analysis.
[DOCX File , 21 KB - medinform_v10i8e37862_app1.docx]

References

1. Karami A, Dahl AA, Turner-McGrievy G, Kharrazi H, Shaw G. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. Intl J Inform Manag 2018 Feb;38(1):1-6. [doi: 10.1016/j.ijinfomgt.2017.08.002]
2. Okuhara T, Ishikawa H, Okada M, Kato M, Kiuchi T. Contents of Japanese pro- and anti-HPV vaccination websites: a text mining analysis. Patient Educ Couns 2018 Mar;101(3):406-413. [doi: 10.1016/j.pec.2017.09.014] [Medline: 29031425]
3. Chen L, Wang X, Peng T. Nature and diffusion of gynecologic cancer-related misinformation on social media: analysis of tweets. J Med Internet Res 2018 Oct 16;20(10):e11515 [FREE Full text] [doi: 10.2196/11515] [Medline: 30327289]
4. Gage-Bouchard EA, LaValley S, Warunek M, Beaupin LK, Mollica M. Is cancer information exchanged on social media scientifically accurate? J Cancer Educ 2018 Dec;33(6):1328-1332 [FREE Full text] [doi: 10.1007/s13187-017-1254-z] [Medline: 28721645]
5. Hornik R, Binns S, Emery S, Epstein VM, Jeong M, Kim K, et al. The effects of tobacco coverage in the public communication environment on young people's decisions to smoke combustible cigarettes. J Commun 2022 Apr;72(2):187-213 [FREE Full text] [doi: 10.1093/joc/jqab052] [Medline: 35386823]
6. Chou WS, Oh A, Klein WM. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
7. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. Health Info Libr J 2017 Dec;34(4):268-283 [FREE Full text] [doi: 10.1111/hir.12192] [Medline: 29045011]
8. Hornik R. Measuring campaign message exposure and public communication environment exposure: some implications of the distinction in the context of social media. Commun Methods Meas 2016 Apr 20;10(2-3):167-169 [FREE Full text] [doi: 10.1080/19312458.2016.1150976] [Medline: 27766123]
9. Shim M, Kelly B, Hornik R. Cancer information scanning and seeking behavior is associated with knowledge, lifestyle choices, and screening. J Health Commun 2006;11 Suppl 1:157-172. [doi: 10.1080/10810730600637475] [Medline: 16641081]
10. Beguerisse-Díaz M, McLennan AK, Garduño-Hernández G, Barahona M, Ulijaszek SJ. The 'who' and 'what' of #diabetes on Twitter. Digit Health 2017 Jan;3:2055207616688841 [FREE Full text] [doi: 10.1177/2055207616688841] [Medline: 29942579]
11. Loeb S, Sengupta S, Butaney M, Macaluso JN, Czarniecki SW, Robbins R, et al. Dissemination of misinformative and biased information about prostate cancer on YouTube. Eur Urol 2019 Apr;75(4):564-567. [doi: 10.1016/j.eururo.2018.10.056] [Medline: 30502104]
12. Park S, Oh H, Park G, Suh B, Bae WK, Kim JW, et al. The source and credibility of colorectal cancer information on Twitter. Medicine (Baltimore) 2016 Feb;95(7):e2775 [FREE Full text] [doi: 10.1097/MD.0000000000002775] [Medline: 26886625]
13. Park MS, He Z, Chen Z, Oh S, Bian J. Consumers' Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites. JMIR Med Inform 2016 Nov 24;4(4):e41. [doi: 10.2196/medinform.5748] [Medline: 27884812]
14. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods Inf Med 2018 Feb 07;41(04):289-298. [doi: 10.1055/s-0038-1634490]
15. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24-29 [FREE Full text] [doi: 10.1197/jamia.M1761] [Medline: 16221948]
16. Doing-Harris KM, Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. J Med Internet Res 2011 May 17;13(2):e37 [FREE Full text] [doi: 10.2196/jmir.1636] [Medline: 21586386]
17. Gu G, Zhang X, Zhu X, Jian Z, Chen K, Wen D, et al. Development of a consumer health vocabulary by mining health forum texts based on word embedding: semiautomatic approach. JMIR Med Inform 2019 May 23;7(2):e12704 [FREE Full text] [doi: 10.2196/12704] [Medline: 31124461]

18. Ibrahim M, Gauch S, Salman O, Alqahtani M. An automated method to enrich consumer health vocabularies using GloVe word embeddings and an auxiliary lexical resource. Peer J Comput Sci 2021;7:e668 [FREE Full text] [doi: 10.7717/peerj-cs.668] [Medline: 34458573]
19. Lazard AJ, Saffer AJ, Wilcox GB, Chung AD, Mackert MS, Bernhardt JM. E-cigarette social media messages: a text mining analysis of marketing and consumer conversations on Twitter. JMIR Public Health Surveill 2016 Dec 12;2(2):e171 [FREE Full text] [doi: 10.2196/publichealth.6551] [Medline: 27956376]
20. Ma T, Atkin D. User generated content and credibility evaluation of online health information: a meta analytic study. Telemat Inform 2017 Aug;34(5):472-486. [doi: 10.1016/j.tele.2016.09.009]
21. Lee K, Hasan S, Farri O, Choudhary A, Agrawal A. Medical concept normalization for online user-generated texts. In: Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI). 2017 Presented at: IEEE International Conference on Healthcare Informatics (ICHI); Aug 23-26, 2017; Park City, UT, USA. [doi: 10.1109/ichi.2017.59]
22. Chunara R, Wisk LE, Weitzman ER. Denominator issues for personally generated data in population health monitoring. Am J Prev Med 2017 Apr;52(4):549-553 [FREE Full text] [doi: 10.1016/j.amepre.2016.10.038] [Medline: 28012811]
23. Plovnick RM, Zeng QT. Reformulation of consumer health queries with professional terminology: a pilot study. J Med Internet Res 2004 Sep 03;6(3):e27 [FREE Full text] [doi: 10.2196/jmir.6.3.e27] [Medline: 15471753]
24. Relia K, Li Z, Cook S, Chunara R. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 U.S. cities. arXiv. Preprint posted online January 31, 2019 [FREE Full text]
25. Fage-Butler AM, Nisbeth Jensen M. Medical terminology in online patient-patient communication: evidence of high health literacy? Health Expect 2016 Jun;19(3):643-653 [FREE Full text] [doi: 10.1111/hex.12395] [Medline: 26287945]
26. Kilbridge KL, Fraser G, Krahn M, Nelson EM, Conaway M, Bashore R, et al. Lack of comprehension of common prostate cancer terms in an underserved population. J Clin Oncol 2009 Apr 20;27(12):2015-2021. [doi: 10.1200/jco.2008.17.3468]
27. van Deursen AJ, van der Zeeuw A, de Boer P, Jansen G, van Rompay T. Digital inequalities in the internet of things: differences in attitudes, material access, skills, and usage. Inform Commun Soc 2019 Jul 27;24(2):258-276. [doi: 10.1080/1369118x.2019.1646777]
28. Din HN, McDaniels-Davidson C, Nodora J, Madanat H. Profiles of a health information-seeking population and the current digital divide: cross-sectional analysis of the 2015-2016 California health interview survey. J Med Internet Res 2019 May 14;21(5):e11931 [FREE Full text] [doi: 10.2196/11931] [Medline: 31094350]
29. Raghavan V, Bollmann P, Jung GS. A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans Inf Syst 1989 Jul;7(3):205-229. [doi: 10.1145/65943.65945]
30. Aidoo M, Harpham T. The explanatory models of mental health amongst low-income women and health care practitioners in Lusaka, Zambia. Health Policy Plan 2001 Jun;16(2):206-213. [doi: 10.1093/heapol/16.2.206] [Medline: 11358923]
31. Mill JE. Describing an explanatory model of HIV illness among aboriginal women. Holist Nurs Pract 2000 Oct;15(1):42-56. [doi: 10.1097/00004650-200010000-00007] [Medline: 12119618]
32. Soffer M, Cohen M, Azaiza F. The role of explanatory models of breast cancer in breast cancer prevention behaviors among Arab-Israeli physicians and laywomen. Prim Health Care Res Dev 2020 Nov 03;21:e48. [doi: 10.1017/s1463423620000237]
33. Zhang Y. Proceedings of the 1st ACM International Health Informatics Symposium. Contextualizing Consumer Health Information Searching: An Analysis of Questions in a Social Q&A Community. In: Proceedings of the 1st ACM International Health Informatics Symposium. 2010 Presented at: IHI '10: ACM International Health Informatics Symposium; Nov 11 - 12, 2010; Arlington Virginia USA. [doi: 10.1145/1882992.1883023]
34. Rovetta A, Castaldo L. A new infodemiological approach through Google Trends: longitudinal analysis of COVID-19 scientific and infodemic names in Italy. BMC Med Res Methodol 2022 Jan 30;22(1):33 [FREE Full text] [doi: 10.1186/s12874-022-01523-x] [Medline: 35094682]
35. Ogden J, Flanagan Z. Beliefs about the causes and solutions to obesity: a comparison of GPs and lay people. Patient Educ Couns 2008 Apr;71(1):72-78. [doi: 10.1016/j.pec.2007.11.022] [Medline: 18201860]
36. Ogden J, Bandara I, Cohen H, Farmer D, Hardie J, Minas H, et al. General practitioners' and patients' models of obesity: whose problem is it? Patient Educ Couns 2001 Sep;44(3):227-233. [doi: 10.1016/s0738-3991(00)00192-0]
37. Institute of Medicine, Board on Neuroscience and Behavioral Health, Committee on Health Literacy. Health Literacy A Prescription to End Confusion. Washington, D.C., United States: National Academies Press; 2004.
38. Niederdeppe J, Levy AG. Fatalistic beliefs about cancer prevention and three prevention behaviors. Cancer Epidemiol Biomarkers Prev 2007 May 01;16(5):998-1003. [doi: 10.1158/1055-9965.EPI-06-0608] [Medline: 17507628]
39. Wang C, Miller SM, Egleston BL, Hay JL, Weinberg DS. Beliefs about the causes of breast and colorectal cancer among women in the general population. Cancer Causes Control 2010 Jan 29;21(1):99-107 [FREE Full text] [doi: 10.1007/s10552-009-9439-3] [Medline: 19787437]
40. Wardle J. Awareness of risk factors for cancer among British adults. Public Health 2001 May;115(3):173-174. [doi: 10.1016/s0033-3506(01)00439-5]
41. Jensen JD, Moriarty CM, Hurley RJ, Stryker JE. Making sense of cancer news coverage trends: a comparison of three comprehensive content analyses. J Health Commun 2010 Mar;15(2):136-151. [doi: 10.1080/10810730903528025] [Medline: 20390983]

42. Brandt AM. Inventing conflicts of interest: a history of tobacco industry tactics. Am J Public Health 2012 Jan;102(1):63-71. [doi: 10.2105/ajph.2011.300292]
43. Petticrew M, Maani Hessari N, Knai C, Weiderpass E. How alcohol industry organisations mislead the public about alcohol and cancer. Drug Alcohol Rev 2018 Mar 07;37(3):293-303. [doi: 10.1111/dar.12596] [Medline: 28881410]
44. Margolin DB. Computational contributions: a symbiotic approach to integrating big, observational data studies into the communication field. Commun Methods Meas 2019 Jul 05;13(4):229-247. [doi: 10.1080/19312458.2019.1639144]
45. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021 Apr 7. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ [accessed 2021-11-03]
46. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: a systematic review. Health Informatics J 2015 Sep;21(3):173-194 [FREE Full text] [doi: 10.1177/1460458213512220] [Medline: 24670899]
47. Fat MJ, Doja A, Barrowman N, Sell E. YouTube videos as a teaching tool and patient resource for infantile spasms. J Child Neurol 2011 Jul 06;26(7):804-809. [doi: 10.1177/0883073811402345] [Medline: 21551373]
48. Liu D, Schuchard H, Burston B, Yamashita T, Albert S. Interventions to reduce healthcare disparities in cancer screening among minority adults: a systematic review. J Racial Ethn Health Disparities 2021 Feb 15;8(1):107-126. [doi: 10.1007/s40615-020-00763-1] [Medline: 32415578]
49. US Preventive Services Task Force, Davidson K, Barry MJ, Mangione CM, Cabana M, Caughey AB, et al. Screening for colorectal cancer: US preventive services task force recommendation statement. JAMA 2021 May 18;325(19):1965-1977. [doi: 10.1001/jama.2021.6238] [Medline: 34003218]
50. Search: list. YouTube Data API. URL: https://developers.google.com/youtube/v3/docs/search/list [accessed 2021-08-10]
51. Rieder B. YouTube data tools. Digital Methods. URL: https://tools.digitalmethods.net/netvizz/youtube/ [accessed 2021-08-10]
52. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013). 2013 Presented at: Advances in Neural Information Processing Systems 26 (NIPS 2013); Dec 5-10, 2013; Lake Tahoe, Nevada, USA.
53. Distributed representations of words using word2vec. GitHub. URL: https://github.com/bnosac/word2vec [accessed 2021-11-03]
54. 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]
55. Kamba M, Manabe M, Wakamiya S, Yada S, Aramaki E, Odani S, et al. Medical needs extraction for breast cancer patients from question and answer services: natural language processing-based approach. JMIR Cancer 2021 Oct 28;7(4):e32005 [FREE Full text] [doi: 10.2196/32005] [Medline: 34709187]
56. Kalyan KS, Sangeetha S. BertMCN: mapping colloquial phrases to standard medical concepts using BERT and highway network. Artif Intell Med 2021 Feb;112:102008. [doi: 10.1016/j.artmed.2021.102008] [Medline: 33581833]
57. Subramanyam KK, S S. Deep contextualized medical concept normalization in social media text. Procedia Comput Sci 2020;171:1353-1362. [doi: 10.1016/j.procs.2020.04.145]
58. Tan AS, Bigman CA. Misinformation about commercial tobacco products on social media—implications and research opportunities for reducing tobacco-related health disparities. Am J Public Health 2020 Oct;110(S3):S281-S283. [doi: 10.2105/ajph.2020.305910]
59. O'Connor A. Coca-cola funds scientists who shift blame for obesity away from bad diets. NY Times. 2015 Aug 9. URL: https://well.blogs.nytimes.com/2015/08/09/coca-cola-funds-scientists-who-shift-blame-for-obesity-away-from-bad-diets/ [accessed 2022-03-07]

Abbreviations

API: application programming interface
PCE: public communication environment
Q&A: question and answer
RQ: research question

Edited by T Hao; submitted 09.03.22; peer-reviewed by C Giraud-Carrier, M Bardus, A Zain; comments to author 04.05.22; revised version received 13.06.22; accepted 22.07.22; published 30.08.22.

Please cite as:
Tong C, Margolin D, Chunara R, Niederdeppe J, Taylor T, Dunbar N, King AJ
Search Term Identification Methods for Computational Health Communication: Word Embedding and Network Approach for Health Content on YouTube
JMIR Med Inform 2022;10(8):e37862
URL: https://medinform.jmir.org/2022/8/e37862
doi:10.2196/37862
PMID:36040760

Original Paper

Multicenter Validation of Natural Language Processing Algorithms for the Detection of Common Data Elements in Operative Notes for Total Hip Arthroplasty: Algorithm Development and Validation

Peijin Han[1], MBBS, MHS; Sunyang Fu[2], PhD; Julie Kolis[3], BS; Richard Hughes[3], PhD; Brian R Hallstrom[3], MD; Martha Carvour[4], MD, PhD; Hilal Maradit-Kremers[2,5], MSc, MD; Sunghwan Sohn[2], PhD; VG Vinod Vydiswaran[6,7], PhD

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States

[2]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, United States

[3]Department of Orthopedic Surgery, University of Michigan, Ann Arbor, MI, United States

[4]Department of Internal Medicine and Epidemiology, University of Iowa, Iowa City, IA, United States

[5]Departments of Orthopedic Surgery, Mayo Clinic, Rochester, MN, United States

[6]Department of Learning Health Sciences, Medical School, University of Michigan, Ann Arbor, MI, United States

[7]School of Information, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
VG Vinod Vydiswaran, PhD
Department of Learning Health Sciences
Medical School
University of Michigan
1161F NIB, 300 N Ingalls St
Ann Arbor, MI, 48109
United States
Phone: 1 734 647 1207
Fax: 1 734 647 3914
Email: vgvinodv@umich.edu

Abstract

Background: Natural language processing (NLP) methods are powerful tools for extracting and analyzing critical information from free-text data. MedTaggerIE, an open-source NLP pipeline for information extraction based on text patterns, has been widely used in the annotation of clinical notes. A rule-based system, MedTagger-total hip arthroplasty (THA), developed based on MedTaggerIE, was previously shown to correctly identify the surgical *approach*, *fixation*, and *bearing surface* from the THA operative notes at Mayo Clinic.

Objective: This study aimed to assess the implementability, usability, and portability of MedTagger-THA at two external institutions, Michigan Medicine and the University of Iowa, and provide lessons learned for best practices.

Methods: We conducted iterative test-apply-refinement processes with three involved sites—the development site (Mayo Clinic) and two deployment sites (Michigan Medicine and the University of Iowa). Mayo Clinic was the primary NLP development site, with the THA registry as the gold standard. The activities at the two deployment sites included the extraction of the operative notes, gold standard development (Michigan: registry data; Iowa: manual chart review), the refinement of NLP algorithms on training data, and the evaluation of test data. Error analyses were conducted to understand language variations across sites. To further assess the model specificity for *approach* and *fixation*, we applied the refined MedTagger-THA to arthroscopic hip procedures and periacetabular osteotomy cases, as neither of these operative notes should contain any *approach* or *fixation* keywords.

Results: MedTagger-THA algorithms were implemented and refined independently for both sites. At Michigan, the study comprised THA-related notes for 2569 patient-date pairs. Before model refinement, MedTagger-THA algorithms demonstrated excellent accuracy for *approach* (96.6%, 95% CI 94.6%-97.9%) and *fixation* (95.7%, 95% CI 92.4%-97.6%). These results were comparable with internal accuracy at the development site (99.2% for *approach* and 90.7% for *fixation*). Model refinement improved accuracies slightly for both *approach* (99%, 95% CI 97.6%-99.6%) and *fixation* (98%, 95% CI 95.3%-99.3%). The

specificity of *approach* identification was 88.9% for arthroscopy cases, and the specificity of *fixation* identification was 100% for both periacetabular osteotomy and arthroscopy cases. At the Iowa site, the study comprised an overall data set of 100 operative notes (50 training notes and 50 test notes). MedTagger-THA algorithms achieved moderate-high performance on the training data. After model refinement, the model achieved high performance for *approach* (100%, 95% CI 91.3%-100%), *fixation* (98%, 95% CI 88.3%-100%), and *bearing surface* (92%, 95% CI 80.5%-97.3%).

Conclusions: High performance across centers was achieved for the MedTagger-THA algorithms, demonstrating that they were sufficiently implementable, usable, and portable to different deployment sites. This study provided important lessons learned during the model deployment and validation processes, and it can serve as a reference for transferring rule-based electronic health record models.

(JMIR Med Inform 2022;10(8):e38155) doi:10.2196/38155

KEYWORDS

total hip arthroplasty; natural language processing; information extraction; model transferability

Introduction

Background

Natural language processing (NLP) methods are powerful tools for extracting information from textual data and are widely applied in medical informatics research [1]. NLP approaches transform unstructured free-text clinical notes into a structured and codified format, thereby reducing human effort on chart reviews in large population-based studies [2-5]. Previous studies have demonstrated that NLP can be an alternative to manual abstraction in many applications, including deidentification, classification, and extraction of medical concepts (eg, clinical symptoms, diagnoses, and medications), semantic modifiers (eg, negation and severity), and temporality information (eg, present vs past; [6,7]). In addition, high-quality NLP approaches applied to real-world data can facilitate clinical registry participation and analysis [8] to further advance clinical research, policy, and surveillance efforts [6,9,10].

In prior research, Wyles et al [11] developed an NLP system to extract common data elements related to total hip arthroplasty (THA) from the operative notes in electronic health records (EHRs). This NLP system contains 3 separate algorithms aimed at capturing the operative *approach*, *fixation* method, and *bearing surface* categories [11,12]. The infrastructure of the NLP system was an open-source NLP pipeline, MedTaggerIE [13], which was developed using an open-source unstructured information management architecture–based information extraction framework [14]. MedTaggerIE contains the following three components: keyword lists (ie, domain-based keywords and short phrases, including wildcard regular expressions), classification rules (ie, regular expression-based patterns to derive the predicted label), and normalization (eg, a standardized form of any THA-related clinical concept). The classification rules take ≥1 regular expression as the input value to extract relevant information. The extracted concepts are normalized to the expected targets as output values. As keywords and phrases containing clinical information can be directly defined by subject matter experts (eg, orthopedic surgeons), the pipeline separates task-specific NLP knowledge engineering from the generic-domain NLP. The final system (referred to as *MedTagger-THA*) was evaluated on 250 THA procedures performed at the Mayo Clinic and demonstrated high accuracy in identifying the abovementioned 3 data elements [11]. The authors found MedTagger-THA to be a promising alternative to the current gold standard of manual chart review for identifying common data elements from orthopedic operative notes [11].

Although typically, the transferability of informatics tools across sites is poor [15] unless explicitly designed for, this data element extraction task is inherently portable across different sites. This is because the development site and the deployment sites (1) share common keywords for *approach* and *fixation* and (2) have common rules to classify *approach* and *fixation*. Some examples of such common rules include labeling “cement femur” and “uncemented shell” as “hybrid” and no “cement” mentions to indicate “uncemented.” However, prior studies have not broadly evaluated whether existing systems, when applied across multiple institutions with heterogeneous EHR systems, are sufficiently implementable (ie, whether the system can be deployed at a different site), usable (ie, whether the system can be easily modified and refined by local users), and portable (ie, whether the system can achieve sufficiently similar results after refinement). Prior studies have shown that significant effort is required for users to apply existing NLP systems [16]. In the context of multi-institutional collaboration, studies have indicated various administrative and implementation challenges such as data privacy; workforce expertise; and the maturity of location extract, transform, and load (ETL) processes [17]. For example, clinical NLP algorithms are often difficult to assess in different hospital settings because of patient confidentiality and difficulties in technology transfer [18]. In addition, the performances of clinical NLP systems, as well as clinical practice and workflows, often vary across institutions and source data [19,20], which results in differences in documentation styles in EHRs [21]. The clinical note structures and languages used within notes can be very different across institutions because of both syntactic variation and semantic variation in the text [21], highlighting the importance of correctly identifying sections [21,22] and semantic lexicon construction for extracting and encoding clinical information from EHRs to achieve semantic interoperability in developing NLP systems [23]. Therefore, to achieve better portability, all these factors must be considered when applying an NLP algorithm developed from one institution to another. In most cases, customization is necessary to achieve a desirable performance and further improve portability.

Objectives

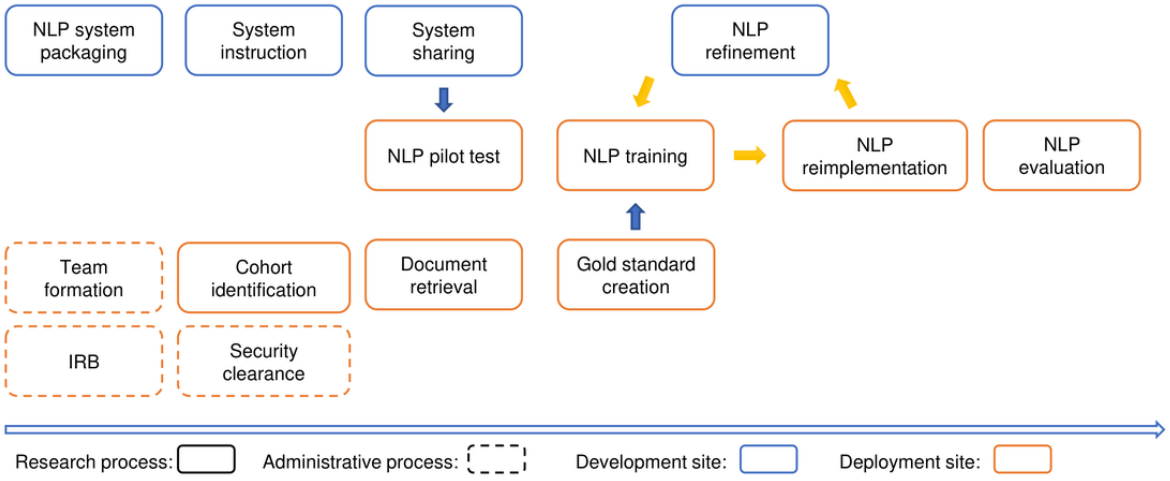
To assess and improve the implementability, portability, and usability of MedTagger-THA, we performed a pilot study to establish an efficient pipeline for transferring MedTagger-THA to 2 external institutions (Michigan Medicine and the University of Iowa) to provide lessons learned for best practices. This study included both common generic processes (eg, task definition, exchanging NLP resources, and training and evaluation) and site-specific processes. Specifically, we established the infrastructure to run MedTagger-THA, including accessing the electronic surgical notes, security clearance for implementation of the MedTagger software tool kit, and running and refining MedTagger-THA. MedTagger-THA algorithms were implemented and refined independently for both sites. At Michigan, we evaluated whether MedTagger-THA can accurately extract information on surgical *approach* and *fixation* from operative notes using the Michigan Arthroplasty Registry Collaborative Quality Initiative (MARCQI) registry as the gold standard. We assessed the out-of-box (prerefinement) validation performances and postrefinement performances on the extraction of *approach* and *fixation*. Finally, we assessed the specificity of these 2 data elements' extraction using periacetabular osteotomy (PAO) and hip arthroscopy cases. As there was no existing arthroplasty registry at the Iowa site, manual chart review was used as the gold standard. We conducted a standardized gold standard development process, which included retrieving operative notes, developing annotation guidelines, and performing corpus annotation. We then used the gold standard to refine and evaluate the MedTagger-THA system for all three data elements—surgical *approach*, *fixation*, and *bearing surface*.

Methods

System Deployment of MedTagger

MedTagger deployment was an iterative test-apply-refinement process involving close collaboration among sites (Figure 1). There were three involved sites: a development site (the site that developed the initial MedTagger-THA system, Mayo Clinic, shown in blue boxes) and 2 deployment sites (Michigan Medicine and the University of Iowa, shown in orange boxes). The initial step was to form an interdisciplinary study team with diverse backgrounds and expertise in orthopedics, information technology, informatics, and epidemiology. Once the team was established, the process was kicked off with several important administrative activities, including institutional review board (IRB) approval and system security clearance.

Figure 1. Overview of the NLP deployment and evaluation process. IRB: institutional review board; NLP: natural language processing.



In addition to the administrative process, research activities were initiated simultaneously. System preparation and packaging were the initial steps at the development site. These steps focused on ascertaining whether the system was usable and interoperable at the deployment site. The NLP system contained two components: (1) a generic MedTagger framework (eg, sentence annotator, tokenizer, and part-of-speech tagger) and (2) MedTagger-THA algorithms (keyword lists and classification rules) that were developed and distributed separately from the main program. This architecture design allows THA algorithms to be easily plugged into the main program for better customizability. Therefore, the initial process was to separate the MedTagger-THA algorithms from the main program in MedTagger for distribution purposes. Following that, the next steps were to prepare the deployment site instructions, which included specifying the input text format (eg, rtf, xml, or plain text), preprocessing instructions, system directories, and system-level instructions and requirements: (1) operating system compatibility (PC, MAC, and Linux), (2) software and packages (Java 1.8), and (3) license (Apache version 2.0). Finally, for code exchange, we used the software development and version control platform Git.

Michigan Site Process

Overview

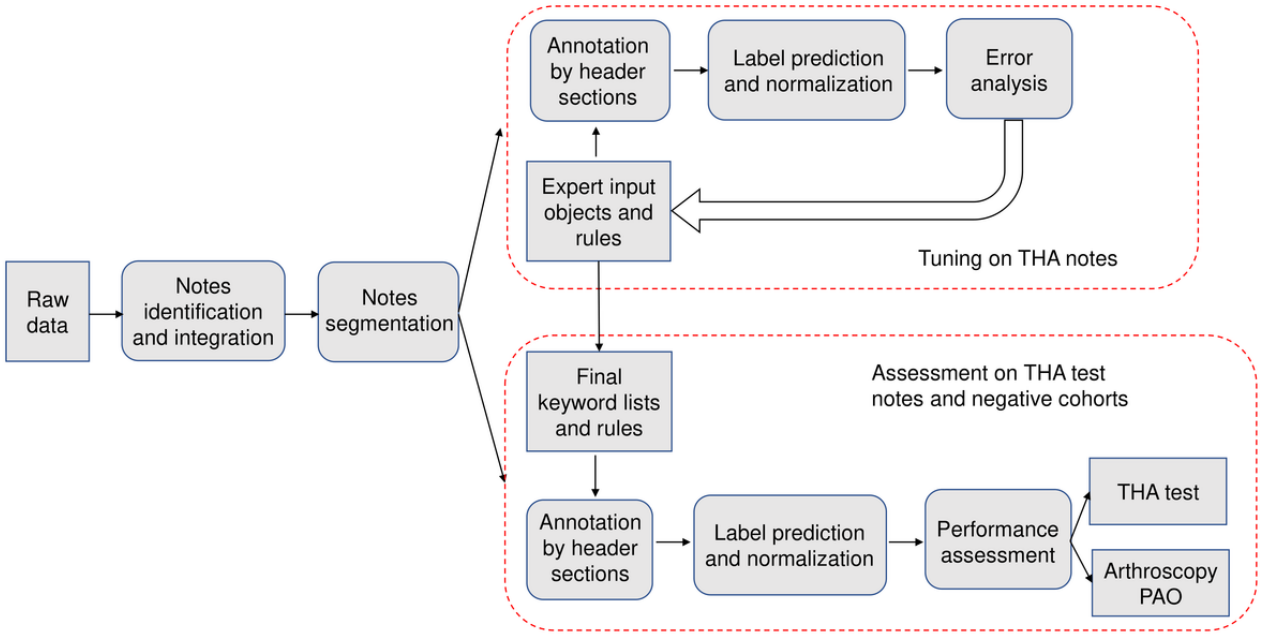
The MARCQI is a group of orthopedic surgeons and medical professionals dedicated to improving the quality of care for patients undergoing hip and knee replacement procedures at Michigan Medicine. The consortium improves the quality of care by addressing variations in patient outcomes related to hip and knee joint replacement surgery [24]. THA cases were abstracted at Michigan Medicine and entered into the MARCQI data repository, including the date of surgery; laterality (left or right); and surgical *approach*, *fixation*, and *bearing surface*. In this study, the MARCQI registry was considered the gold standard to evaluate the automated algorithms. The surgical *approach* documented in the MARCQI included "Anterior," "Anterolateral," "Posterior," and "Transtrochanteric." The *fixation* methods included "Cemented," "Uncemented," "Hybrid," and "Reverse Hybrid." The *bearing surface* materials included "Ceramic-on-polyethylene," "Metal-on-polyethylene," and "Dual Mobility."

We extracted the operative notes for elective and conversion primary THA performed between January 1, 2014, and April 30, 2019, from the Epic-based Michigan Medicine EHR system. As the *bearing surface* was captured by catalog numbers of implants used and not by notes abstraction, we only assessed the accuracy, precision, recall, and F_1-score of the algorithms on *approach* and *fixation*. All 95% CIs were obtained using the procedure by Agresti and Coull [25].

In addition to THA, PAO and arthroscopy procedures are also conducted in Michigan Medicine and are sometimes applied to patients with THA. As these surgical procedures have some common features (such as *approach*), we believe it is necessary to assess the specificity of the algorithm to evaluate whether it is overly generalized. To assess the specificity of *fixation*, we applied the algorithms to PAO and hip arthroscopy cases as neither of these should have any kind of fixation that we were assessing. Hip arthroscopy cases were also used to assess the specificity of the algorithms for identifying the *approach* as arthroscopic hip procedures should not have an identified *approach*, as they were conducted through portals.

The *note-processing pipeline* that we established involved several steps (Figure 2).

Figure 2. The workflow of the note-processing pipeline at the Michigan site. The rectangles represent the data and the rounded rectangles represent the process. PAO: periacetabular osteotomy; THA: total hip arthroplasty.



Notes Identification and Integration

We first identified distinct patient-date pairs from THA notes, which represented procedures conducted on certain dates over specific individuals. For each patient, we ordered the notes by note documentation time and gathered all the notes that were within a 15-day interval as a note set for 1 operation. For 1 note set, we took the first documentation time to represent the patient's procedure date. We then mapped patient-date pairs to the MARCQI data set. For patients with PAO and arthroscopy, we used the same 15-day window to integrate notes for unique patient-date pairs.

Notes Segmentation

For each unique patient-date pair, we first segmented the note sets by section headers. The section headers parsed from the THA notes are listed in Table S1 in Multimedia Appendix 1, which include concepts of preoperative diagnosis, procedure, findings, and implants. Among these headers, the section headers that were most likely to be semantically related to "procedures" (Table S2 in Multimedia Appendix 1) were predefined in the

Michigan data. To refine the MedTagger-THA model using Michigan data, we first randomly split the data set into training (80%) and test (20%) sets based on unique patients. As the MARCQI only began to collect *fixation* data in 2017, THA notes before 2017 were excluded from these analyses.

Annotation by Header Sections

For each unique patient-date pair, the *approach* and *fixation* keywords were extracted from all relevant sections. The initial *approach* and *fixation* keywords were predefined using the keyword lists published previously [11]. As defined in the study by Wyles et al [11], "The assertion of each concept includes certainty (i.e., positive, negative, and possible) along with the person who experienced the event (i.e., the patient or someone else, such as husband, child, etc.), whereas temporality identifies the timing of an event (i.e., historical or present)." Concept with "positive" certainty, "present" temporality, and the "patient" who experienced the event is the concept of interest.

Label Prediction and Normalization

Classification rules comprising regular expressions were applied to derive prediction labels. The initial classification rules have been published previously [11]. For *approach*, the labels included "Anterior," "Anterolateral," "Posterior," and "Transtrochanteric." For *fixation*, the labels included "Cemented," "Hybrid," "Uncemented," and "Reverse Hybrid." The prediction labels also included two special conditions—if no annotation was given by any section, the final prediction would be "missing," and if multiple annotations were given but were not the same, the final prediction would be "ambiguous." For both the training and test sets, we applied MedTagger-THA [11] to extract the *approach* and *fixation* and evaluated their out-of-box performance.

Error Analysis

We then worked with the MARCQI abstraction professional to resolve the misclassifications, missing predictions, and ambiguous predictions in the training data set. We iteratively tuned the MedTagger-THA model [26] by adding keywords to the *approach* and *fixation* keyword lists and modifying the classification rules until the model performance could not be improved on the training data set. The test data set was not used during the refining process. After the refining process, we obtained the updated keyword lists and classification rules (Table S3 in Multimedia Appendix 1). Thus, in the following text, the refined MedTagger-THA obtained is referred to as MedTagger-THA-Michigan.

Assessment of THA Test Notes

We assessed the performance of MedTagger-THA-Michigan on the test data set. We further performed an error analysis on the test data set to analyze the limitations of the model. Finally, we evaluated the specificity of *approach* and *fixation* extraction from PAO and hip arthroscopy cases. Figure 2 shows the workflow of the Michigan identification pipeline.

Iowa Site Process

We concurrently deployed the system at the University of Iowa. The gold standard corpus for the evaluation of the NLP system was established through a standard corpus annotation process [27]. A trained nurse abstractor manually reviewed 100 operative reports randomly sampled from known THA procedures between January 1, 2009, and December 31, 2016, from Iowa's Epic-based EHRs. Questions regarding the abstracted data were resolved upon consultation with a physician with content expertise. Chart review was conducted using the same concept definition as that based on the total joint arthroplasty registry; in addition to *approach* and *fixation*, data collection included *bearing surface* classified into four categories: metal-on-polyethylene, ceramic-on-polyethylene, metal-on-metal, and ceramic-on-ceramic. The gold standard data set was equally split into 2 subsets of 50 training instances and 50 test instances. We followed an iterative training and refining process [26] to evaluate and refine the NLP algorithms. Briefly, the prototype system, MedTagger-THA, was applied to the training data. Error cases were manually reviewed by a team of researchers at Iowa with experience in informatics and clinical documentation to identify key errors or themes leading to missing or misclassified results. The keywords were manually curated through an iterative refining process until all major issues were resolved.

Ethics Approval

The study was approved by the IRBs at both the University of Michigan (HUM00143841) and the University of Iowa (201903205).

Results

Michigan Site Results

For THA notes, 2304 unique patients with 2569 patient-date pairs were mapped to the MARCQI registry data set. From the PAO notes and arthroscopy notes, 398 and 523 patient-date pairs were extracted, respectively. For *approach* and *fixation*, the out-of-box external validation of the MedTagger-THA algorithms demonstrated excellent accuracy (surgical *approach*: 96.6%, 95% CI 94.6%-97.9%; *fixation*: 95.7%, 95% CI 92.4%-97.6%; Tables 1 and 2).

Table 1. Out-of-box performance of MedTagger-total hip arthroplasty (THA) for surgical approach: comparison of the gold standard (registry data) and notes classified by MedTagger-THA in the training and test data.[a]

Gold standard	MedTagger-THA, n (%)				
	Anterior	Anterolateral	Posterior	Ambiguous	Missing inference
Training data (n=2062)					
Anterior	261 (12.7)	0 (0)	2 (0.1)	1 (0)	0 (0)
Anterolateral	0 (0)	1 (0)	2 (0.1)	0 (0)	1 (0)
Posterior	4 (0.2)	2 (0.1)	1737 (84.2)	1 (0)	50 (2.4)
Test data (n=507)					
Anterior	68 (13.4)	0 (0)	0 (0)	0 (0)	0 (0)
Anterolateral	0 (0)	1 (0.2)	0 (0)	0 (0)	0 (0)
Posterior	0 (0)	1 (0.2)	421 (83)	0 (0)	15 (3)
Transtrochanteric	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.2)

[a]Accuracy: 96.6% (95% CI 94.6%-97.9%); precision: 99.8% (95% CI 98.7%-100%); recall: 96.6% (95% CI 94.6%-97.9%); F_1-score: 98.2% (95% CI 96.5%-99.1%).

Table 2. Out-of-box performance of MedTagger-total hip arthroplasty (THA) for fixation: comparison of the gold standard (registry data) and notes classified by MedTagger-THA in the training and test data.[a]

Gold standard	MedTagger-THA, n (%)			
	Cemented	Hybrid	Uncemented	Ambiguous
Training data (n=1053)				
Cemented	0 (0)	1 (0.1)	0 (0)	0 (0)
Hybrid	1 (0.1)	76 (7.2)	3 (0.3)	17 (1.6)
Uncemented	0 (0)	29 (2.8)	925 (87.8)	1 (0.1)
Test data (n=256)				
Cemented	0 (0)	0 (0)	0 (0)	0 (0)
Hybrid	0 (0)	23 (9)	2 (0.8)	5 (2)
Uncemented	0 (0)	4 (1.6)	222 (86.7)	0 (0)

[a]Accuracy: 95.7% (95% CI 92.4%-97.6%); precision: 95.7% (95% CI 92.4%-97.6%); recall: 95.7% (95% CI 92.4%-97.6%); F_1-score: 95.7% (95% CI 92.4%-97.6%).

The classification errors, ambiguous cases, and missing inferences are listed in Table 3. Classification errors for *approach* occurred when (1) the notes in one section contained mentions for a different *approach*, whereas the mentions for the correct *approach* were missing; (2) the mentions for a different *approach* were extracted from sections other than "procedure and findings"; and (3) the section of "procedure and findings" contained many different mentions for *approach*. Ambiguous cases occurred when mentions for the correct *approach* were extracted from notes related to "procedures and findings," and different *approach* mentions were also extracted from other sections for a single surgery. Missing inferences occurred when the mentions for *approach* were missing in the notes or when the mentions were misspelled. Common classification errors for *fixation* occurred when the certainty of inference was incorrectly assessed. For example, for "non cemented stem," the certainty was assessed as "positive" instead of "negative," which resulted in an "Uncemented" *fixation* instance being misclassified as "Hybrid." If the stem mentioned in the notes was not included in the predefined keyword list (eg, "femur"), a "Hybrid" instance was misclassified as "Uncemented," or a "Cemented" instance was misclassified as "Hybrid." "Hybrid" instances could also be misclassified as "Cemented" when "Cemented" was explicitly stated in the notes and a *Stem Concept* was noted, as the algorithm treated "Cemented" as a direct mention of *cemented fixation*. Similar situations were observed in ambiguous cases, where some sections misclassified "Hybrid" instances as "Cemented," whereas others gave the correct classification. An "Uncemented" instance was inferred as a default *fixation* label when there was no mention of the "cement concept." Therefore, if there was no mention of the "cement concept" explicitly, even if the surgery was "Cemented" or "Hybrid," it was classified as "Uncemented."

Table 3. Classification errors and ambiguous cases for approach and fixation in the Michigan data set.

Keyword	Classification error	Ambiguous cases	Missing
Approach	• The mention of the correct *approach* was missing, although the mentions for other approaches existed. • The notes in the "Complications" section contained mentions for a different *approach*, whereas the mentions for the correct *approach* were missing. • Multiple different *approach* mentions were extracted from the same section, and the *approach* mentions that appeared more times were given priority.	• Notes related to diagnosis sections but not the procedures contained different mentions of *approach*; for example, "Left hip osteoarthritis, abductor deficiency (sclerosed greater trochanter with chronic avulsion of gluteus medius)" was annotated as "anterolateral," but the gold standard label was "posterior." • Notes related to "indications" contained hypothetical conditions; for example, "We offered her the option of anterior or posterior *approach* and she decided that an anterior *approach* was preferable." was annotated as "posterior" instead of "anterior."	• Direct mentions of *approach* were not included in the keyword list; for example, "posterior THA[a] precautions," "APPROACH: Posterior," and "posterolateral." • No mentions indicating the *approach* as the notes referred to previous incisions. • Misspelling of the mentions led to unrecognition (eg, "shortrotators").
Fixation	• "Uncemented" was misclassified as "Hybrid" The note mentioned "non cement stem" but the certainty of the inference was positive for the *Cement Concept*.[b] • "Hybrid" was misclassified as "Uncemented"; for example, "femur" was not included in the stem keyword list, and no *Cement Concept* was mentioned in the notes. The surgeries were "Total Hip Replacement with Computer Navigation." • "Cemented" was misclassified as "Hybrid" as "femur" was not included in the stem keyword list, *Shell Concept*[b] was also excluded. Only *Cement Concept* led to "Hybrid"; for example, "A polyethylene acetabular liner was cemented in using the trabecular metal acetabular revision system longevity, 0-degree face angle, 36-millimeter inner diameter VerSys Hip prosthesis standard neck offset size 11 was cemented into the femur." • "Hybrid" was misclassified as "Cemented" as "Cemented" was a direct mention and had priority over others; for example: "Total Hip Arthroplasty, cemented, Right Hip" was misclassified as "Cemented" In the notes, only the femoral canal is cemented.	• For a single surgery note, some sections misclassified "Hybrid" as "Cemented" as "Cemented" was a direct mention of *Cement Concept* and had the highest priority over others; for example, "Total Hip Arthroplasty, cemented femoral stem" was misclassified as "cemented" instead of "Hybrid."	• Missingness in *fixation* was set to "Uncemented."

[a]THA: total hip arthroplasty.

[b]Concept name.

After model refinement (Tables 4 and 5), the validation accuracies improved for both surgical *approach* and *fixation* (*approach*: 99%, 95% CI 97.6%-99.6% vs 96.6%; *fixation*: 98%, 95% CI 95.3%-99.3% vs 95.7%). Giving priorities to sections related to "procedures" reduced the ambiguous cases for *fixation* (from 5 to 2). For specificity assessment, we identified the *approach* mentioned in 11.1% (58/523) of patient-date pairs for the arthroscopy data set (specificity: 465/523, 88.9%). These false positives were mainly because of the keywords for the approach mentioned in the notes, such as "Hana table," "anterior superior iliac spine," or "tensor fascia lata," although these mentions described positioning and portal placement. At times, arthroscopy was combined with PAO in a procedure, and the mentions for *approach* could be related to PAO. We did not identify any *fixation* mentioned in the PAO cohort or in the arthroscopy cohort (specificity 100%).

Table 4. *Approach* after refinement: comparison of the gold standard and notes classified by refined MedTagger-total hip arthroplasty (THA) in the Michigan test data set (N=507).[a]

Gold standard	MedTagger-THA-Michigan, n (%)				
	Anterior	Anterolateral	Posterior	Ambiguous	Missing inference
Anterior	68 (13.4)	0 (0)	0 (0)	0 (0)	0 (0)
Anterolateral	0 (0)	1 (0.2)	0 (0)	0 (0)	0 (0)
Posterior	0 (0)	0 (0)	434 (85.6)	0 (0)	3 (0.6)
Transtrochanteric	0 (0)	0 (0)	1 (0.2)	0 (0)	0 (0)

[a]Accuracy: 99% (95% CI 97.6%-99.6%); precision: 99.6% (95% CI 98.4%-100%); recall: 99% (95% CI 97.6%-99.6%); F_1-score: 99.3% (95% CI 98%-99.8%).

Table 5. *Fixation* after refinement: comparison of the gold standard and notes classified by refined MedTagger-total hip arthroplasty (THA) in the Michigan test data set (N=256).[a]

Gold standard	MedTagger-THA-Michigan, n (%)			
	Cemented	Hybrid	Uncemented	Ambiguous
Cemented	0 (0)	0 (0)	0 (0)	0 (0)
Hybrid	1 (0.4)	26 (10.2)	1 (0.4)	2 (0.8)
Uncemented	0 (0)	1 (0.4)	225 (87.9)	0 (0)

[a]Accuracy: 98% (95% CI 95.3%-99.3%); precision: 98% (95% CI 95.3%-99.3%); recall: 98% (95% CI 95.3%-99.3%); F_1-score: 98% (95% CI 95.3%-99.3%).

Iowa Site Results

No registry data were available at the University of Iowa. Therefore, we performed a manual chart review of a total of 100 operative reports (50 training reports and 50 test reports) and tested the performance of MedTagger-THA on this data set for *approach* (Table 6), *fixation* (Table 7), and *bearing surface* (Table 8). Overall, the model achieved moderate-high performance on the training data, with the lowest performance observed for the *bearing surface* concept. Model refinement included modifying the default output for the *bearing surface* to match the case distribution of Iowa's data and adding additional *liner*-related concepts (eg, A-class liner) to improve the sensitivity of the *fixation* category. After model refinement, the model achieved high performance for all three data elements: *approach* (100%, 95% CI 91.3%-100%), *fixation* (98%, 95% CI 88.3%-100%), and *bearing surface* (92%, 95% CI 80.5%-97.3%).

Table 6. *Approach*: comparison of the gold standard and notes classified by MedTagger-total hip arthroplasty (THA) in the University of Iowa data set (N=100).[a]

Gold standard	MedTagger-THA-Iowa, n (%)			Total, n (%)
	Anterior	Anterolateral	Posterior	
Training data (n=50)				
Anterior	12 (24)	1 (2)	0 (0)	13 (26)
Anterolateral	0 (0)	0 (0)	0 (0)	0 (0)
Posterior	0 (0)	0 (0)	37 (74)	37 (74)
Test data (n=50)				
Anterior	14 (28)	0 (0)	0 (0)	14 (28)
Anterolateral	0 (0)	0 (0)	0 (0)	0 (0)
Posterior	0 (0)	0 (0)	36 (72)	36 (72)

[a]Accuracy: 100% (95% CI 91.3%-100%); precision 100% (95% CI 91.3%-100%); recall: 100% (95% CI 91.3%-100%); F_1-score: 100% (95% CI 91.3%-100%).

Table 7. *Fixation*: comparison of the gold standard and notes classified by MedTagger-total hip arthroplasty (THA) in the University of Iowa data set (N=100).[a]

Gold standard	MedTagger-THA-Iowa, n (%)			Total, n (%)
	Cemented	Hybrid	Uncemented	
Training data (n=50)				
Cemented	0 (0)	0 (0)	0 (0)	0 (0)
Hybrid	0 (0)	1 (2)	0 (0)	1 (2)
Uncemented	0 (0)	0 (0)	49 (98)	49 (98)
Test data (n=50)				
Cemented	0 (0)	0 (0)	0 (0)	0 (0)
Hybrid	1 (2)	0 (0)	0 (0)	1 (2)
Uncemented	0 (0)	0 (0)	49 (98)	49 (98)

[a]Accuracy: 98% (95% CI 88.3%-100%); precision: 98% (95% CI 88.3%-100%); recall: 98% (95% CI 88.3%-100%); F_1-score: 98% (95% CI 88.3%-100%).

Table 8. *Bearing surface*: comparison of the gold standard and notes classified by MedTagger-total hip arthroplasty (THA) in the University of Iowa data set (N=100).[a]

Gold standard	MedTagger-THA-Iowa, n (%)				Total, n (%)
	MoP[b]	CoP[c]	MoM[d]	CoC[e]	
Training data (n=50)					
MoP	25 (50)	1 (2)	1 (2)	0 (0)	27 (54)
CoP	0 (0)	17 (34)	0 (0)	0 (0)	17 (34)
MoM	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CoC	0 (0)	6 (12)	0 (0)	0 (0)	6 (12)
Test data (n=50)					
MoP	20 (40)	2 (4)	0 (0)	0 (0)	22 (44)
CoP	0 (0)	26 (52)	0 (0)	0 (0)	26 (52)
MoM	0 (0)	0 (0)	0 (0)	1 (2)	1 (2)
CoC	0 (0)	1 (2)	0 (0)	0 (0)	1 (2)

[a]Accuracy: 92% (95% CI: 80.5%-97.3%); precision: 92% (95% CI 80.5%-97.3%); recall: 92% (95% CI 80.5%-97.3%); F_1-score: 92% (95% CI 80.5%-97.3%).

[b]MoP: metal-on-polyethylene

[c]CoP: ceramic-on-polyethylene.

[d]MoM: metal-on-metal.

[e]CoC: ceramic-on-ceramic.

Discussion

Principal Findings

In this study, we applied the MedTagger-THA algorithms developed at Mayo Clinic to the THA operative notes at Michigan Medicine and the University of Iowa. The algorithms were implementable, usable, and portable, with high performances at both deployment sites. Model refinements for major or recurring errors further improved the accuracy. In NLP reimplementation studies, refinement of the original model to "adapt" to the local health care system is important for the portability of the EHR models. We plan to validate MedTagger-THA in different hospital settings and EHRs and integrate these adapted models back into the original model. We expect that the continuous model refinement will further enhance portability.

We learned many important lessons from the NLP deployment and evaluation across different institutions. When assessing implementability, we encountered several workforce-related, institutional policy–related, and data infrastructure–related challenges and gaps. First, successful deployment and evaluation require at least three types of expertise: orthopedic domain knowledge of total joint arthroplasty, ETL skills, and expertise in NLP and model evaluation. We observed variable expertise at different sites and a strong need for multidisciplinary team science collaboration. Second, institutional policies have a significant impact on the time and effort related to the exchange of informatics resources. For example, the process of obtaining

security clearances for sharing NLP systems to a locally secured environment could range from days to months depending on institutional policies. We also discovered a variation of strictness among institutions for sharing the NLP results for error analysis and refinement, suggesting the need for early planning and communication for multisite NLP research beyond just a multi-institutional IRB. The third aspect is the maturity of ETL and data infrastructure. There is substantial variation in institutional ETL processes and personnel training because of different data infrastructures. An institution with lower data infrastructure maturity would involve a manual abstraction process as an alternative, which can be a huge barrier for high-throughput NLP solutions. Specifically, the data infrastructure at Mayo Clinic is a centralized unified data platform, a duplication of the Epic Clarity table for handling various data retrieval requests in a central location. In contrast, Iowa has several decentralized enterprise data warehouses that require multiple ETL processes for data retrieval. Michigan maintains a separate research data warehouse for clinical and translational research, with a separate ETL pipeline to populate the warehouse with structured and free-text data. The aforementioned findings indicate the high complexity and dynamics of the multi-institutional EHR environment and suggest the need for a situated contextual understanding of multisite clinical NLP research.

When assessing usability and portability, there are some caveats in the process of NLP model refinement. We noticed that giving priorities to sections that related to "procedures" reduced the ambiguous cases. The headers of these sections may vary from site to site and require curation by medical experts to guarantee semantic interoperability. It is always possible to add curated keywords to the keyword list; however, these keywords may not be compatible with the original settings. For example, the negation algorithm was adopted from *ConText* [28]. "Posterior THA precautions" and "posterior THA" were considered "negated" in the original MedTagger-THA algorithms, as "precautions" is an indicator of "possible" instead of "positive" certainty according to *ConText* [28]. However, these mentions were indications of the posterior *approach* in Michigan's data. We also changed the rules for identifying *fixation* better in Michigan's data; however, we were not sure whether these changes would compromise the model performance at Mayo Clinic. These observations indicate the need to differentiate portable components of the model from institution-specific components that do not generalize well across institutions. Therefore, in the future refinement of MedTagger-THA, we suggest that a panel of medical experts and abstraction specialists from both the development site and validation and deployment sites should determine which changes can be incorporated into the original model for further distribution and better portability and which changes should be retained at the local validation site for institution-specific performance improvements.

We also noticed that *approach* and *fixation* were not unique mentions in THA notes. Keywords for the THA approach can be mentioned in other procedures, such as total knee arthroplasty, PAO, and arthroscopy, although those descriptions were not related to THA. As MedTagger-THA extracted information based on keyword mentions and rules defined by a series of regular expressions, we should acknowledge that the model should only be applied to THA notes. Therefore, before applying the MedTagger-THA model, it is necessary to filter out the non-THA operative notes. This process is relatively straightforward using text-based search and filtering, as the procedure names are usually explicitly mentioned in the "procedure" section.

MedTagger-THA algorithms are very useful for identifying THA-related data elements; however, they have several important limitations. MedTagger-THA was developed based on keywords and classification rules. Although we were able to extract keywords mentioned if the misspelled keywords were found during curation and training, future versions of MedTagger-THA should incorporate a validated spell check and correction model. In addition, MedTagger-THA cannot recognize hypothetical alternate treatment plans, such as whether the procedure was actually performed or merely documented as differentially discussed. MedTagger-THA links concepts by their locations in the texts (eg, *Cement Concept* close to *Stem Concept* means the stem is cemented) but cannot process the contextualized information (eg, 2 concepts were not related to each other). To solve these problems, we plan to conduct future research focusing on understanding the contextualized information when performing named entity recognition tasks using more advanced NLP techniques, such as methods based on machine learning, including deep learning models. Finally, for the Iowa site, the data for algorithm validation and refinement may be biased from the Iowa population of patients with THA because of the small sample size (n=100) and only one annotator being involved. Validation and refinement using small sample sizes may be valid in centers where clinical practice variability is low and thus, might increase accessibility to NLP-based tools where data infrastructural resources are limited or in development.

Conclusions

In conclusion, MedTagger-THA algorithms were sufficiently implementable, usable, and portable to different deployment sites for *approach* and *fixation* identification from THA notes. *Bearing surface* identification may be subject to greater variability in clinical practice patterns and surgical devices. As expected, model refinement within unique institutional EHRs is useful for improving accuracy. This study underscores the importance of undertaking such model refinements in institutional settings and informs future implementation efforts to enhance transferability across institutions.

Acknowledgments

The work was supported by Hilal Maradit Kremers' National Institutes of Health grant (R01 AR73147), with Michigan Medicine and the University of Iowa as subaward sites. The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the University of Michigan or the University of Iowa.

The authors would like to acknowledge He Jintao, MS, for his contribution to compiling the natural language processing modules at the Iowa site.

Authors' Contributions

VGVV, SS, HMK, MC, and RH conceived and designed this study. SF and PH developed the models. PH wrote the first draft, and all authors helped with interpreting the results and with the final review of the manuscript.

Conflicts of Interest

BRH's employer receives partial salary support from Blue Cross Blue Shield of Michigan for their work as Co-Director of MARCQI.

Multimedia Appendix 1
Section headers in operative notes, headers related to "procedures" and updated keyword lists, and classification rules for the total hip arthroplasty approach and fixation classification (Michigan).
[DOCX File , 28 KB - medinform_v10i8e38155_app1.docx]

References

1. Koleck T, Dreisbach C, Bourne P, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
2. Wi C, Sohn S, Rolfes MC, Seabright A, Ryu E, Voge G, et al. Application of a natural language processing algorithm to asthma ascertainment. An automated chart review. Am J Respir Crit Care Med 2017 Aug 15;196(4):430-437 [FREE Full text] [doi: 10.1164/rccm.201610-2006OC] [Medline: 28375665]
3. Alzu'bi AA, Watzlaf VJ, Sheridan P. Electronic health record (EHR) abstraction. Perspect Health Inf Manag 2021 Mar 15;18(Spring):1g [FREE Full text] [Medline: 34035788]
4. 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]
5. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005;12(4):448-457 [FREE Full text] [doi: 10.1197/jamia.M1794] [Medline: 15802475]
6. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
7. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: overview and actionable suggestions for future advances. J Biomed Inform 2018 Dec;88:11-19 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.005] [Medline: 30368002]
8. Shah RF, Bini S, Vail T. Data for registry and quality review can be retrospectively collected using natural language processing from unstructured charts of arthroplasty patients. Bone Joint J 2020 Jul;102-B(7_Supple_B):99-104. [doi: 10.1302/0301-620x.102b7.bjj-2019-1574.r1]
9. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
10. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. Int J Med Inform 2019 Sep;129:13-19 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.05.018] [Medline: 31445247]
11. Wyles CC, Tibbo ME, Fu S, Wang Y, Sohn S, Kremers WK, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. J Bone Joint Surg 2019;101(21):1931-1938. [doi: 10.2106/jbjs.19.00071]
12. Fu S, Leung LY, Wang Y, Raulli A, Kallmes DF, Kinsman KA, et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. JMIR Med Inform 2019 Apr 21;7(2):e12109 [FREE Full text] [doi: 10.2196/12109] [Medline: 31066686]
13. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:149-153 [FREE Full text] [Medline: 24303255]

14. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng 1999;10(3-4):327-348. [doi: 10.1017/s1351324904003523]
15. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 6;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
16. Zheng K, Vydiswaran VG, Liu Y, Wang Y, Stubbs A, Uzuner O, et al. Ease of adoption of clinical natural language processing software: an evaluation of five systems. J Biomed Inform 2015 Dec;58 Suppl:S189-S196 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.008] [Medline: 26210361]
17. Liu S, Wen A, Wang L, He H, Fu S, Miller R, National COVID Cohort Collaborative, Natural Language Processing, Subgroup, National COVID Cohort Collaborative (N3C). An open natural language processing development framework for EHR-based clinical research: a case demonstration using the national COVID cohort collaborative (N3C). arXiv 2021 Oct 20 [FREE Full text]
18. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020 Feb;145(2):463-469 [FREE Full text] [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
19. Fan J, Prasad R, Yabut RM, Loomis RM, Zisook DS, Mattison JE, et al. Part-of-speech tagging for clinical text: wall or bridge between institutions? AMIA Annu Symp Proc 2011;2011:382-391 [FREE Full text] [Medline: 22195091]
20. Wagholikar KB, Torii M, Jonnalagadda SR, Liu H. Pooling annotated corpora for clinical concept extraction. J Biomed Semantics 2013 Jan 08;4(1):3 [FREE Full text] [doi: 10.1186/2041-1480-4-3] [Medline: 23294871]
21. Sohn S, Wang Y, Wi C, Krusemark EA, Ryu E, Ali MH, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. J Am Med Inform Assoc 2018 Mar 01;25(3):353-359 [FREE Full text] [doi: 10.1093/jamia/ocx138] [Medline: 29202185]
22. Edinger T, Demner-Fushman D, Cohen AM, Bedrick S, Hersh W. Evaluation of clinical text segmentation to facilitate cohort retrieval. AMIA Annu Symp Proc 2017;2017:660-669 [FREE Full text] [Medline: 29854131]
23. Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K, et al. Towards a semantic lexicon for clinical natural language processing. AMIA Annu Symp Proc 2012;2012:568-576 [FREE Full text] [Medline: 23304329]
24. Michigan Arthroplasty Registry Collaborative Quality Initiative. URL: https://marcqi.org/ [accessed 2022-03-21]
25. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 1998 May;52(2):119-126. [doi: 10.1080/00031305.1998.10480550]
26. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: a methodology review. J Biomed Inform 2020 Sep;109:103526 [FREE Full text] [doi: 10.1016/j.jbi.2020.103526] [Medline: 32768446]
27. Fu S, Leung LY, Raulli A, Kallmes DF, Kinsman KA, Nelson KB, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. BMC Med Inform Decis Mak 2020 Mar 30;20(1):60 [FREE Full text] [doi: 10.1186/s12911-020-1072-9] [Medline: 32228556]
28. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 2009 Oct;42(5):839-851 [FREE Full text] [doi: 10.1016/j.jbi.2009.05.002] [Medline: 19435614]

Abbreviations

EHR: electronic health record
ETL: extract, transform, and load
IRB: institutional review board
MARCQI: Michigan Arthroplasty Registry Collaborative Quality Initiative
NLP: natural language processing
PAO: periacetabular osteotomy
THA: total hip arthroplasty

Edited by T Hao; submitted 21.03.22; peer-reviewed by J Shi, M Torii; comments to author 04.05.22; revised version received 30.05.22; accepted 12.07.22; published 31.08.22.

Please cite as:
Han P, Fu S, Kolis J, Hughes R, Hallstrom BR, Carvour M, Maradit-Kremers H, Sohn S, Vydiswaran VGV
Multicenter Validation of Natural Language Processing Algorithms for the Detection of Common Data Elements in Operative Notes for Total Hip Arthroplasty: Algorithm Development and Validation
JMIR Med Inform 2022;10(8):e38155
URL: https://medinform.jmir.org/2022/8/e38155
doi:10.2196/38155
PMID:36044253

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org