JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 8 (2020), Issue 11 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Viewpoints

Review

Corrigenda and Addenda

Original Paper

Temporal Design Patterns for Digital Phenotype Cohort Selection in Critical Care: Systematic Literature Assessment and Qualitative Synthesis

Daniel Capurro[1,2], MD, PhD; Mario Barbe[3,4], MD, MSc; Claudio Daza[2], BSc, MPH; Josefa Santa Maria[2], MD; Javier Trincado[2], MD

[1]School of Computing and Information Systems, Centre for Digital Transformation of Health, University of Melbourne, Melbourne, Australia

[2]Department of Internal Medicine, School of Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile

[3]Department of Biomedical Informatics, Clinica Alemana, Santiago, Chile

[4]Instituto de Ciencias e Innovación en Medicina, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, Chile

Corresponding Author:
Daniel Capurro, MD, PhD
School of Computing and Information Systems
Centre for Digital Transformation of Health
University of Melbourne
Room 3.24, Level 3, Doug McDonnel (Building 168)
Parkville Campus
Melbourne, 3010
Australia
Phone: 61 8344 4504
Email: dcapurro@unimelb.edu.au

Abstract

Background: Inclusion criteria for observational studies frequently contain temporal entities and relations. The use of digital phenotypes to create cohorts in electronic health record–based observational studies requires rich functionality to capture these temporal entities and relations. However, such functionality is not usually available or requires complex database queries and specialized expertise to build them.

Objective: The purpose of this study is to systematically assess observational studies reported in critical care literature to capture design requirements and functionalities for a graphical temporal abstraction-based digital phenotyping tool.

Methods: We iteratively extracted attributes describing patients, interventions, and clinical outcomes. We qualitatively synthesized studies, identifying all temporal and nontemporal entities and relations.

Results: We extracted data from 28 primary studies and 367 temporal and nontemporal entities. We generated a synthesis of entities, relations, and design patterns.

Conclusions: We report on the observed types of clinical temporal entities and their relations as well as design requirements for a temporal abstraction-based digital phenotyping system. The results can be used to inform the development of such a system.

(JMIR Med Inform 2020;8(11):e6924) doi:10.2196/medinform.6924

KEYWORDS

digital phenotyping; clinical data; temporal abstraction

Introduction

The increasing costs of health care [1] and the rapid advance of new discoveries create the need for streamlining the identification of effective health interventions. The evidence-based clinical practice paradigm promotes the generation of such knowledge through high-quality randomized controlled trials and systematic reviews [2]. However, when we consider the amount of resources required to conduct a randomized controlled trial [3], alternative ways to assess the effectiveness of clinical interventions become attractive.

The broad adoption of electronic health records (EHRs) [4] allows researchers to analyze routinely collected electronic clinical data to conduct comparative effectiveness research. A

health care system that systematically analyzes clinical data to generate and test hypotheses should be able to learn from itself, becoming a learning health care system [5]. However, converting a traditional health care system into a learning one faces several organizational, societal, and data-related barriers.

"Good data" is a relative concept [6], because it depends on who the user is and what the data are being used for. When EHR data are collected primarily for direct patient care and not with the explicit objective of generating knowledge, a majority of the captured information is stored as free text or other types of unstructured format, limiting its reuse potential. Our group has estimated that 75% of all data elements required for calculating clinical quality measures are not available as structured and computable database fields [7]. Similar results have been found about the clinical information required for clinical trial or cohort eligibility criteria [8]. The combination of data and rules to specify the latter are denominated a phenotyping algorithm [9]; digital phenotypes are the cornerstone of generating new knowledge from routinely collected clinical data and of a learning health care system.

The value of structured data lies in its capacity of being computed without major processing, therefore several attempts have been made to overcome the lack of structured clinical data. A review by Shivade et al [10] reported that the most frequently used methods to automatically identify patient cohorts based on EHR phenotypes were rule-based systems, natural language processing, and machine learning techniques. In this review a majority of studies involved the use of diagnostic codes to select eligible patients. However, although wide variations are seen, diagnostic codes frequently present poor sensitivity and specificity to accurately determine patients' conditions [11].

Despite current advances in the area, cohort building systems require a significant amount of effort to develop and test and, in real scenarios, the most commonly used strategy to deal with limited EHR data quality is to use a combination of simple rules and manual verification of clinical data from patient records [12]. Thus, the field is still open to new and complementary approximations to identify patient cohorts based on digital phenotypes.

Clinical researchers face many barriers when querying clinical databases to find patients that match a specific cohort definition. One problem is that querying clinical databases is a complex task requiring multiple interactions between clinical researchers and database experts. Among those complexities, inclusion criteria frequently define temporal patterns of clinical events, which need convoluted temporal database queries [13]. This is needed in up to 40% of studies [8]. Finding patients that meet certain temporal patterns of clinical events could be both a barrier—when systems that do not easily support this feature are not available—and a very powerful tool to accurately retrieve patient cohorts based on these temporal digital phenotypes. However, systems that easily support this feature are not readily available.

In this study, we systematically reviewed the critical care literature to characterize the temporal representation of inclusion criteria, interventions, and outcomes, used by clinical researchers when designing a clinical study. The product of this review is a set of basic temporal entities, temporal relations, and the resulting temporal phenotype design patterns. The results can be used to inform the design of temporal abstraction-based digital phenotyping systems.

Methods

Data Source

We conducted a systematic literature review of published articles in the critical care domain. Using the Web of Science Journal Citation Reports, we selected the top 5 critical care journals according to their impact factor. Paired reviewers (MB, CD, JT, and JS) manually reviewed all publications and decided on inclusion or exclusion according to criteria described in the following section. Disagreements were solved by consensus.

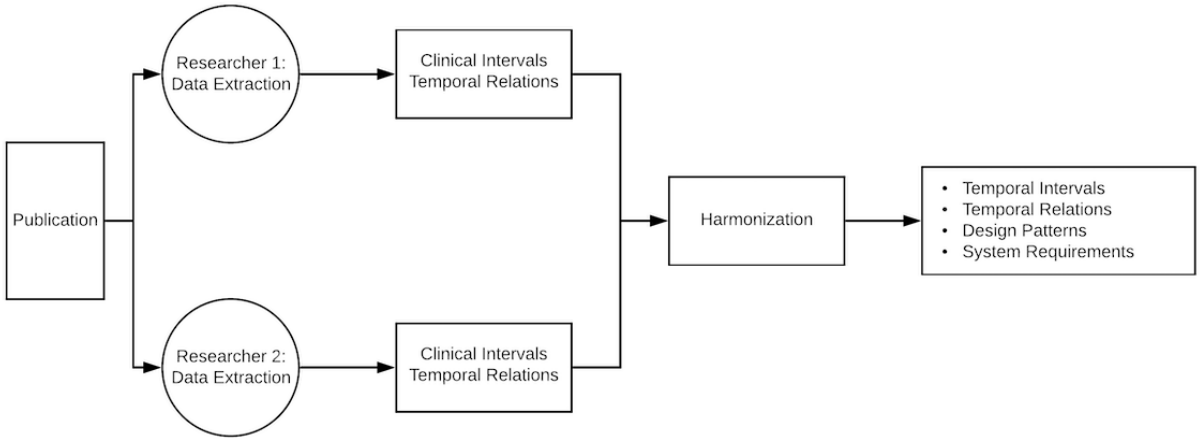
Types of Studies Included

We included retrospective studies conducted in intensive care unit settings which used data obtained from EHRs, clinical databases generated from EHRs, or through manual chart abstractions. We excluded studies which presented exclusively outpatient or emergency department data.

Data Extraction

For every included study paired reviewers (MB, CD, JT, and JS) manually identified and extracted—using a purposefully built online form—all elements characterizing the study's inclusion criteria, the interventions or exposures being studied (or the comparison group), and primary outcomes as defined by the original study authors following the Patient/Population, Intervention, Comparison, Outcome (PICO) framework [14]. Each attribute was then classified according to the clinical type (diagnosis, vital sign, laboratory result, medication, etc). When these elements contained a temporal dimension as defined by Boland et al [15], they were abstracted as *temporal intervals* or *instants*. For example, if the study included patients that underwent mechanical ventilation, because mechanical ventilation occurs during a period of time, such inclusion criteria would be abstracted as a mechanical ventilation interval; in the case of a single dose of antibiotics, that would be abstracted as a drug administration instant. Attributes that were not suitable to be represented as temporal attributes—such as sex, race—were represented as nontemporal patient attributes. A representation of the data extraction process can be seen in Figure 1.

Figure 1. Overview of the data extraction process.



When possible, if an interval or instant was itself an abstraction of lower-level concepts, it was decomposed into its parts according to the description explicitly provided or cited by the authors. If there were no details in the paper, we used standard definitions, when available. For example, when a systemic inflammatory response syndrome [16] was used as an inclusion criterion, we abstracted its components as determined by systemic inflammatory response syndrome definition at the time of the study: body temperature, heart rate, respiratory rate, arterial CO_2 pressure, and white blood cells. If standard definitions were not available, we did not decompose that interval and it was extracted as the authors described it. Clinical events that are stored as free-text format—whether because they are traditionally stored in this form or it is the only available format, such as radiology reports or surgical protocols—were not represented in the abstractions.

To minimize variability in the data extraction process, all researchers followed an initial training period. Researchers classified the identified elements—inclusion criteria, interventions or exposures, and outcomes—using the framework described above. Discrepancies on the concept extraction and temporal representations were resolved by group agreement. We performed descriptive statistics from the concept extractions and the temporal elements obtained.

The abstraction process was conducted iteratively and continued until the point of saturation. We predefined saturation as being met when including additional studies did not add any new types of temporal elements.

Finally, researchers systematically documented temporal and nontemporal relationships between the identified temporal elements. This allowed us to identify the temporal query design patterns present in the literature. Finally, we documented the required functionality for a novel temporal abstraction-based system to identify patient cohorts, interventions/exposures, and outcomes in large clinical databases.

Results

Data Extraction

After iteratively extracting clinical concepts, the point of saturation—where no new types of temporal elements were identified—was reached after reviewing 28 primary studies. We obtained a total of 362 clinical entities, 48.6% (n=176) were inclusion criteria, 24.3% (n=88) were classified as interventions or exposures, and 27.0% (n=98) were outcomes. Abstracted entities were further classified into categories according to their clinical type, which are described, with examples, in Table 1. Therapeutic interventions (26.2%, 95/362), diagnostic tests (20.7%, 75/362), and vital signs (11.3%, 41/362) categories covered almost 60% of all entities.

Table 1. Categories, examples, and frequencies of identified clinical entities (N=362).

Classification	Example	Count, n (%)
Therapeutic intervention	Drugs or procedures: vancomycin, orotracheal intubation	95 (26.2)
Laboratory/diagnostic tests	Serum creatinine, hematocrit	75 (20.7)
Vital signs	Body temperature, respiratory rate, central venous pressure	41 (11.3)
Diagnosis	Pneumonia, urinary infection	35 (9.7)
Patient location	Intensive care unit hospitalization, patient transfer	26 (7.2)
Clinical scores	APACHE II[a], Cerebral Performance Category	25 (6.9)
Nontemporal attribute	Sex, ethnicity	17 (4.7)
Death	In-hospital deaths, 30-day mortality	15 (4.1)
Physical examination finding	Pupil diameter, abdominal pain	11 (3.0)
Past medical history	History of trauma	7 (1.9)
Disposition	Discharge to home, institution, or other health center	5 (1.4)
Other	Appropriate antibiotic usage	10 (2.8)

[a]APACHE II: Acute Physiology And Chronic Health Evaluation II.

Temporal Entities—Instants and Intervals

Of the 362 abstracted entities, 328 could be classified as clinical instants or intervals. Most entities could be abstracted as instants (54.1%, 196/362). This type of abstraction is used to represent a clinical event that does not have a duration but has a timestamp. For example, one inclusion criteria in this category was “the presence of arterial lactate > 2.5 mmol/L.” As much as 36.5% (132/362) of abstracted entities were of type interval. This type of abstraction is used to represent a clinical event that has a duration—defined by a start and end time—greater than 0. An example of a clinical interval is “noninvasive mechanical ventilation for at least 48 hours.”

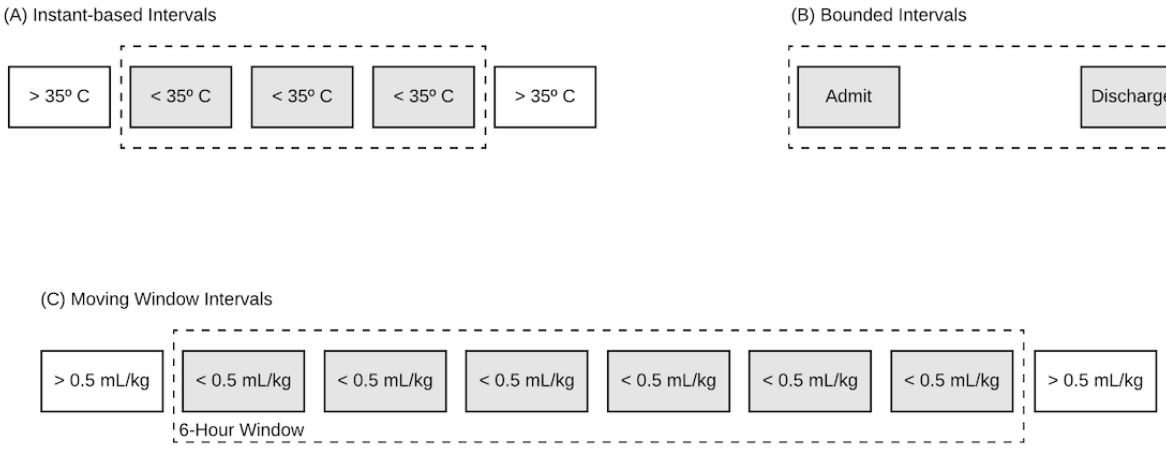
Types of Clinical Intervals

Further analysis of clinical intervals showed that they can also be subdivided into 3 different categories:

- Instant-based intervals: clinical intervals that are abstractions of identical instants. An example of this is hypothermia interval in which the interval is an abstraction of multiple instants of low body temperature measurements. Sometimes specific conditions have to be met to abstract this kind of interval: a time interval for a patient receiving more than 100 mL/hour of intravenous fluids. In other occasions, the instants were only used as categorical variables, regardless of the quantity: patient receiving normal saline infusion.
- Bounded intervals: clinical intervals that are abstractions of specific instants defining their start and end times. An example of this is a hospitalization interval, where the start is defined by an admission instant and the end is defined by a discharge instant. Additional arithmetic operations may need to be applied to these intervals, for example, a clinical interval describing a hospitalization longer than 7 days.
- Moving window intervals: clinical intervals where a specific condition needs to be met during a predefined window of time. An example of this was an oliguria interval, in which the condition oliguria (urinary output < 0.5 mL/kg/hour) has to be met during a 6-hour window. This denomination is consistent with previous descriptions [17].

Graphic examples of the 3 types of intervals are presented in Figure 2.

Figure 2. Three observed categories of clinical temporal intervals.



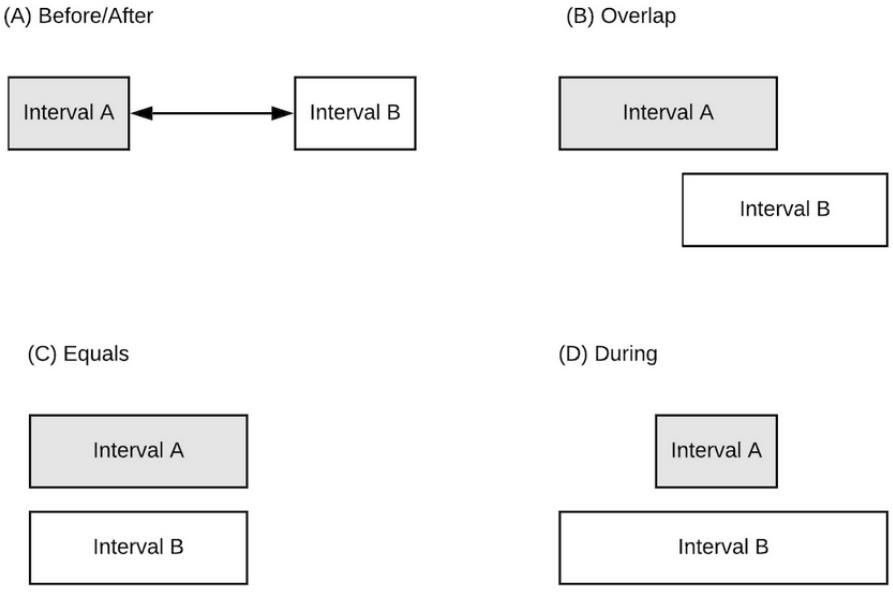
Within-Interval Calculations

In a small subset of intervals, arithmetic calculations were needed to correctly abstract them. For example, calculating an interval of pulse pressure variation (PPV) within a defined range would require calculating PPV (%) = 100 × 2 ([PP_{max} – PP_{min}]/[PP_{max} + PP_{min}]) at each instant before executing the abstraction. Other examples of within-interval calculations included counting the number of instants occurring inside an abstracted interval. An example of this would be an outcome defined as the number of chest x-rays performed on each patient during his or her stay in the intensive care unit; the interval is of type bounded (admission/discharge from the intensive care unit) and we need to count the number of additional instants (chest x-rays) occurring within the interval.

Temporal and Atemporal Relations

We explored the temporal relations between instants and intervals and, as expected, all of them conformed to the temporal logic described by Allen [18]. Briefly, Allen described 13 possible temporal relations between a pair of intervals. Examples of these are the before, equal, and overlap temporal relations, among others. Graphic examples are presented in Figure 3.

Figure 3. Examples of temporal relations.



In addition, some intervals were constructed by combinations of Boolean relations between intervals and instants. For example, to adequately represent a pediatric sepsis interval as defined by the International Pediatric Sepsis Consensus Conference [19] as required by the study authors, we required the Boolean relation AND between 6 different instants, and each one of them temporally related to an instant-based interval.

Some of the extracted entities did not have a temporal component and were denominated nontemporal patient attributes. Examples of these are age, race, and sex.

Finally, 5.5% (20/362) of the extracted concepts were not able to be represented using this proposed framework. For example, the outcome appropriate antimicrobial administration defined as whether the isolated bacteria were susceptible to the

administered antibiotic implies a qualitative interpretation of a laboratory examination, which is out of scope of a temporal representation of clinical entities.

Nested Queries

One additional functionality that was particularly salient was the need to perform nested queries. In a nested query, a query uses the output of another query as its input. Observational studies frequently explore the effect of a specific exposure; this study design involves creating 2 patient cohorts that are identical except for the exposure. When the outcome is assessed in both cohorts, a nested query is the most natural way to satisfy this requirement:

SELECT (Outcome Phenotype) FROM (Exposure Cohort)

In this case *Outcome Phenotype* and *Exposure Cohort* are both themselves queries.

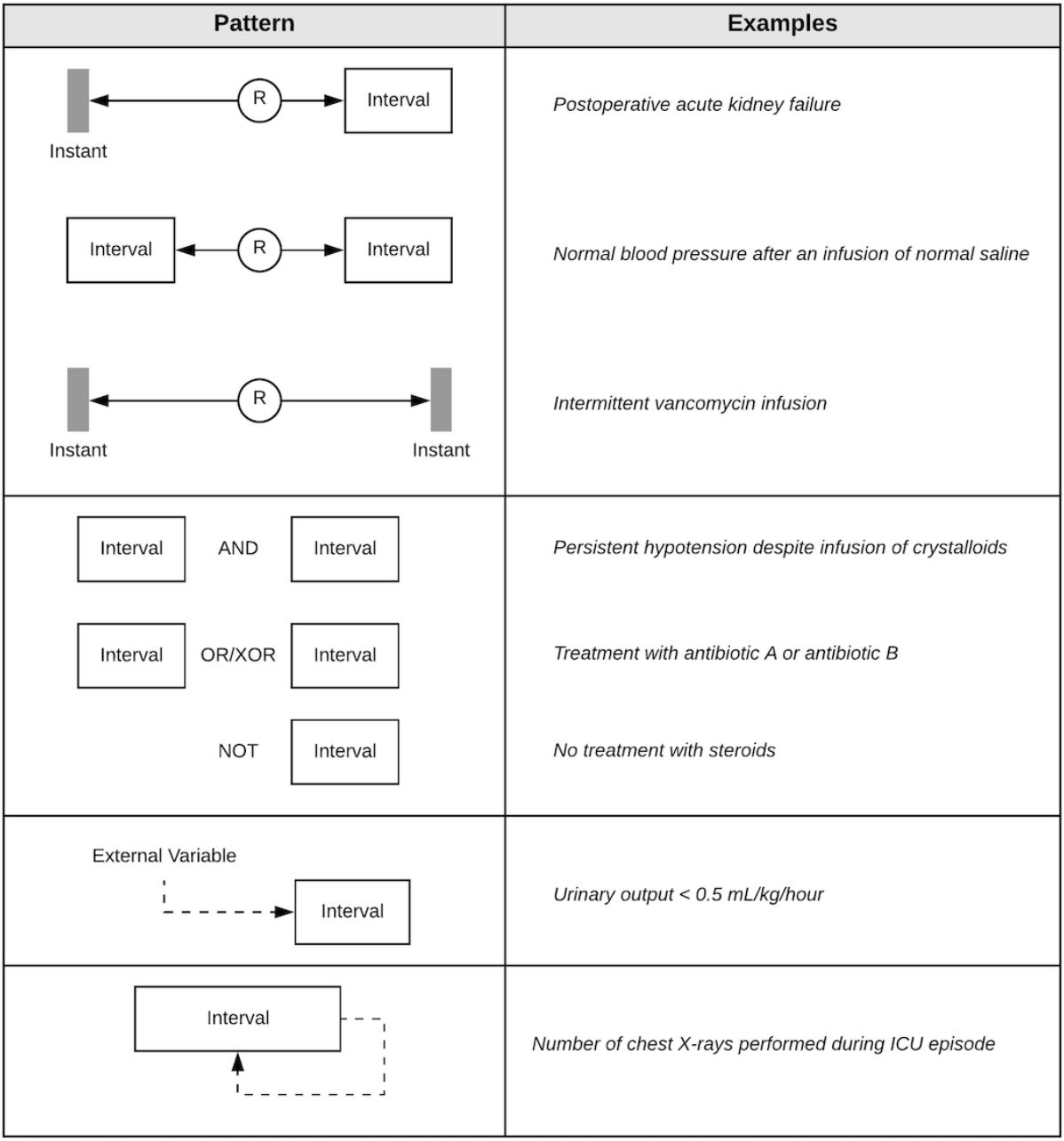
Design Patterns

The combination of temporal entities (instants and intervals), temporal relations, and nontemporal patient attributes can be used to describe the different observed patterns. A graphic description is presented in Figure 4.

Intervals can be temporally related to either instants or intervals. The same was observed for instants. We observed all 13 temporal relations described by Allen [18]. We call a pair of temporally related temporal entities (ie, Interval–Relation–Interval) a basic pattern. These basic patterns can, in turn, be related to other temporal entities or other temporal patterns. Those relations can be either temporal or through Boolean operators (AND, OR, Exclusive OR, NOT).

Intervals can use external variables as a condition to meet either before or after being abstracted. The first case would be the abstraction of an interval of reduced urinary output (<5 mL/kg/hour), in which each urinary output instant needs to be checked against the patient's body weight (the external variable) before being added to the interval. An example for the second case could be *total dose of prednisone less than 10 mg/kg*. This interval is abstracted from individual instants of prednisone administration and after the interval is abstracted, it is checked against the patient's body weight (the external variable). A final case was seen when an internal calculation—using information completely contained within the interval—was needed to be performed to generate the required attribute for the interval. An example of this would be an interval of a series of chest x-rays and, at the end, the number of x-rays would be calculated to create the interval total number of chest x-rays per week.

Figure 4. Examples of identified clinical temporal design patterns. ICU: intensive care unit.



Discussion

Main Findings

This study presents a systematic, literature-based assessment of design requirements to develop a temporal abstraction-based digital phenotyping tool. Such a tool would facilitate the conduction of retrospective clinical studies in critical care using routinely collected electronic clinical data through enabling a rich description of clinical phenotypes. Once validated, these temporally abstracted digital phenotypes should be able to correctly represent patient cohorts, clinical interventions or exposures, as well as relevant clinical outcomes. The iterative nature of this review, which was conducted until reaching information saturation, adds robustness to its findings.

The initial findings of this review are consistent with previous research describing the nature of temporal clinical entities, in the form of clinical instants and intervals [20], as well as temporal relationships between these entities [18]. Other temporal abstraction-based digital phenotyping systems have been described in the past [21,22]; however, there are no reports that their development has been informed by systematically reviewing observational studies. As a consequence, this study adds 3 additional functionalities that may facilitate the creation of digital phenotypes for observational research.

First, this review shows that 3 subtypes of clinical intervals—instant-based, bounded, and moving window—are necessary to adequately represent digital phenotypes. Second, in addition to these interval subtypes, there is a need to perform

calculations both within a clinical interval and with data external to the interval being abstracted. The third component involves the need to allow for nested queries when building digital phenotypes for observational studies. Other findings of this systematic review confirm the need to query for temporal relations and Boolean relations as described by Mo et al [23] in their desiderata for digital phenotyping. Finally, it is essential to highlight the need to generate high-quality temporal metadata during routine clinical documentations because temporal queries are an essential component of digital phenotyping.

Limitations

The main limitation of this review is its focus only on intensive care studies. We chose this setting given the temporal density of clinical data collected during critical care episodes. We cannot claim that these findings will be similar in other clinical domains; that statement would need to be explicitly verified in additional studies. A second limitation is the exclusion of inclusion criteria based on free text contained in clinical notes or reports. This was an explicit decision given our goal of designing a digital phenotyping system able to abstract higher-level concepts from structured data without relying on free text. We still need to demonstrate the feasibility of this approach [24].

Acknowledgments

This work was supported by DC's CONICYT-FONDECYT (Chile) grant (no. 11130577).

Authors' Contributions

MB contributed with data analysis, manuscript write-up, and final document review. CD, JS, and JT contributed with data analysis and final document review. DC contributed with study conceptualization and design, data analysis, manuscript write-up, final document review, and decision to submit.

Conflicts of Interest

None declared.

References

1. Schäferhoff M, Martinez S, Ogbuoji O, Sabin ML, Yamey G. Trends in global health financing. BMJ 2019 May 20;365:l2185 [FREE Full text] [doi: 10.1136/bmj.l2185] [Medline: 31109918]
2. Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target Validity and the Hierarchy of Study Designs. Am J Epidemiol 2019 Feb 01;188(2):438-443 [FREE Full text] [doi: 10.1093/aje/kwy228] [Medline: 30299451]
3. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003 Sep 24;290(12):1624-1632. [doi: 10.1001/jama.290.12.1624] [Medline: 14506122]
4. Palabindala V, Pamarthy A, Jonnalagadda NR. Adoption of electronic health records and barriers. J Community Hosp Intern Med Perspect 2016;6(5):32643 [FREE Full text] [doi: 10.3402/jchimp.v6.32643] [Medline: 27802857]
5. Budrionis A, Bellika JG. The Learning Healthcare System: Where are we now? A systematic review. J Biomed Inform 2016 Dec;64:87-92. [doi: 10.1016/j.jbi.2016.09.018] [Medline: 27693565]
6. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC) 2016;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
7. Capurro D, Yetisgen M, van Eaton E, Black R, Tarczy-Hornoch P. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. EGEMS (Wash DC) 2014;2(1):1079 [FREE Full text] [doi: 10.13063/2327-9214.1079] [Medline: 25848594]
8. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. Summit Transl Bioinform 2010 Mar 01;2010:46-50 [FREE Full text] [Medline: 21347148]
9. Overby CL, Weng C, Haerian K, Perotte A, Friedman C, Hripcsak G. Evaluation considerations for EHR-based phenotyping algorithms: A case study for drug-induced liver injury. AMIA Jt Summits Transl Sci Proc 2013;2013:130-134 [FREE Full text] [Medline: 24303321]
10. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
11. Khan A, Ramsey K, Ballard C, Armstrong E, Burchill LJ, Menashe V, et al. Limited Accuracy of Administrative Data for the Identification and Classification of Adult Congenital Heart Disease. J Am Heart Assoc 2018 Jan 12;7(2):e007378 [FREE Full text] [doi: 10.1161/JAHA.117.007378] [Medline: 29330259]
12. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform 2010 Jun;43(3):451-467 [FREE Full text] [doi: 10.1016/j.jbi.2009.12.004] [Medline: 20034594]

13. Hruby GW, Matsoukas K, Cimino JJ, Weng C. Facilitating biomedical researchers' interrogation of electronic health record data: Ideas from outside of biomedical informatics. J Biomed Inform 2016 Apr;60:376-384 [FREE Full text] [doi: 10.1016/j.jbi.2016.03.004] [Medline: 26972838]
14. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc 2006:359-363 [FREE Full text] [Medline: 17238363]
15. Boland MR, Tu SW, Carini S, Sim I, Weng C. EliXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria. AMIA Jt Summits Transl Sci Proc 2012;2012:71-80 [FREE Full text] [Medline: 22779055]
16. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. Chest 1992 Jun;101(6):1644-1655. [doi: 10.1378/chest.101.6.1644] [Medline: 1303622]
17. Gall W, Duftschmid G, Dorda W. Moving time window aggregates over patient histories. Int J Med Inform 2001 Oct;63(3):133-145. [doi: 10.1016/s1386-5056(01)00164-2] [Medline: 11502429]
18. Allen J. Maintaining knowledge about temporal intervals. Commun. ACM 1983 Nov;26(11):832-843 [FREE Full text] [doi: 10.1145/182.358434]
19. Goldstein B, Giroir B, Randolph A, International Consensus Conference on Pediatric Sepsis. International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. Pediatr Crit Care Med 2005 Jan;6(1):2-8. [doi: 10.1097/01.PCC.0000149131.72248.E6] [Medline: 15636651]
20. Shahar Y. A framework for knowledge-based temporal abstraction. Artificial Intelligence 1997 Feb;90(1-2):79-133 [FREE Full text] [doi: 10.1016/s0004-3702(96)00025-2]
21. Post AR, Kurc T, Willard R, Rathod H, Mansour M, Pai AK, et al. Temporal abstraction-based clinical phenotyping with Eureka!. AMIA Annu Symp Proc 2013;2013:1160-1169 [FREE Full text] [Medline: 24551400]
22. Mate S, Bürkle T, Kapsner LA, Toddenroth D, Kampf MO, Sedlmayr M, et al. A method for the graphical modeling of relative temporal constraints. J Biomed Inform 2019 Dec;100:103314. [doi: 10.1016/j.jbi.2019.103314] [Medline: 31629921]
23. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc 2015 Nov;22(6):1220-1230 [FREE Full text] [doi: 10.1093/jamia/ocv112] [Medline: 26342218]
24. Hernandez-Boussard T, Monda KL, Crespo BC, Riskin D. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. J Am Med Inform Assoc 2019 Nov 01;26(11):1189-1194 [FREE Full text] [doi: 10.1093/jamia/ocz119] [Medline: 31414700]

Abbreviations

APACHE II: Acute Physiology And Chronic Health Evaluation II
EHR: electronic health record
PICO: Patient/Population, Intervention, Comparison, Outcome
PPV: pulse pressure variation

Edited by G Eysenbach; submitted 09.07.20; peer-reviewed by L Jordan, G Jiang; comments to author 17.08.20; revised version received 30.08.20; accepted 28.10.20; published 24.11.20.

Please cite as:
Capurro D, Barbe M, Daza C, Santa Maria J, Trincado J
Temporal Design Patterns for Digital Phenotype Cohort Selection in Critical Care: Systematic Literature Assessment and Qualitative Synthesis
JMIR Med Inform 2020;8(11):e6924
URL: http://medinform.jmir.org/2020/11/e6924/
doi:10.2196/medinform.6924
PMID:33231554

Viewpoint

Assessment of mHealth Interventions: Need for New Studies, Methods, and Guidelines for Study Designs

Roxana Ologeanu-Taddei[1], PhD

TBS Business School, Toulouse, France

Corresponding Author:
Roxana Ologeanu-Taddei, PhD
TBS Business School
1, place Alphonse Jourdain
Toulouse, 31068
France
Phone: 33 5 61 29 48 51
Email: r.ologeanu-taddei@tbs-education.fr

Abstract

This viewpoint argues that the clinical effects of mobile health (mHealth) interventions depends on the acceptance and adoption of these interventions and their mediators, such as usability of the mHealth software, software performance and features, training and motivation of patients and health care professionals to participate in the experience, or characteristics of the intervention (eg, personalized feedback).

(JMIR Med Inform 2020;8(11):e21874) doi:10.2196/21874

KEYWORDS

eHealth; mHealth; usability; management; survey; trust; guidelines; evaluation

Background

In the past years, designs of research studies in medical journals have been formalized according to the reporting guidelines of academic associations, international consortia, and publishers, which enable publications of clinical trials, systematic reviews, and meta-analyses. The use of mobile health (mHealth) technologies is expected to increase, creating new paths for health care delivery. However, there are no specific guidelines to enable researchers to design and present their studies and results on this topic, except for the existing guidelines on randomized clinical trials (RCTs), which can be used for certain assessments of mHealth interventions.

Measures of Clinical Effects Require Opening the Black Box of Information Technology

There is no doubt that RCTs are useful to assess the clinical outcomes and effectiveness of mHealth. In this context, regulatory bodies such as the US Food and Drug Administration (FDA) [1] and European regulations have recently updated the requirements for clinical proofs of mHealth solutions. Nevertheless, the focus on RCT methods leads to the black boxing of mHealth interventions [2], which means that the technology is considered as a homogeneous device or as a pharmaceutical substance. This view misses the main characteristic of an mHealth intervention (and overall, that of information technology [IT])—the embeddedness of data and clinical processes (as reflected in the guidelines for diagnosis and personalized monitoring). Therefore, assessing the clinical effect of an mHealth intervention should disentangle the effect of a new process of personalized monitoring, the effect of the ubiquitous access enabled by mobile devices, and the comprehension and adoption of clinical guidelines implemented into the application. In addition, mHealth solutions may differ from one another because of their different designs of these processes.

Need of Standard Guidelines to Assess Technology and Mediators of Outcomes

Moreover, the clinical effect of mHealth solutions depends on the acceptance and adoption of these solutions and their mediators, such as usability of the mHealth software, software performance and features, training and motivation of patients and health care professionals to participate in the experience, or characteristics of the intervention (eg, personalized feedback). For example, a clinical effect such as survival benefits for patients with cancer who use a surveillance mHealth app depends on the acceptance and adoption of the app, which can

be influenced by the usability of the app and patients' prior experience in using mobile apps, motivation or trust in IT, or other alternative mediators contributing to the main reported outcome; this influences are often neglected by RCTs.

Settings of mHealth interventions must be carefully described and assessed in hypothesis-generating studies [3], such as observational studies and case reports. These studies can identify specific moderators and mediators that state for whom and under what conditions the health intervention works [3]. Moderators may identify population groups with possible causal mechanisms or courses of illness. The mHealth mediators identify possible causal mechanisms, meaning causal links between the intervention and the outcome, through which the intervention may achieve its effects [3]. As a next step, these moderators and mediators should be considered as stratification variables in forthcoming RCTs focused on hypothesis testing. Otherwise, RCTs are likely to be based on weak assumptions rather than empirical evidence.

Beyond Effectiveness: Risk Assessment

In addition, RCTs need to be complemented by other clinical trials and case reports to assess safety risks [4] and unintended consequences of mHealth. The acknowledgment of these risks is at the core of the updated regulations for medical devices, which include software and mHealth. However, in most cases, assessments of mHealth safety risks are conducted separately [5], for example, in feasibility or usability studies, which use different methods of varying rigor and do not generate cumulative knowledge. In addition, case reports on the adoption stages of mHealth solutions should be inspired by engineering methods (eg, fault tree analysis rather than pharmacovigilance studies). Moreover, relevant and cumulative knowledge can be gathered if publications on the issues of usability and user acceptance are presented not only in health informatics journals but also in major medical journals, as these issues cause clinical effects. For example, the Journal of the American Medical Association (JAMA) Network advises authors to use the guidelines of the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network [6], which include guidelines related to "economic studies." Similar initiatives should be undertaken for studies concerning mHealth. In recent years, cumulative knowledge has been gathered on the risks associated with poor usability of health IT [7]. In line with this literature, a step forward was taken only for mHealth solutions, which are qualified as medical devices [8]. However, even for those applications, national and international regulations (ie, CE marking in Europe or FDA regulations) and harmonized standards (ie, EN 62366 advised for CE marking in Europe) strengthened the requirements for premarket certifications but did not standardize a threshold for usability or technical performance. We must recognize that recommending the minimum required sample size (eg, 15 users identified by user profile numbers) makes an improvement to the summative usability assessment method [9]. It is also necessary to assess user profiles accurately. These user profiles may be related to health care occupations (eg, clinical secretaries, physicians, or nurses), different health departments (eg, infectious diseases or pediatrics), and social or demographic variables. In addition, several other characteristics (eg, computer literacy, prior experience of using mobile apps, and users' engagement or trust in IT) may mediate or moderate mHealth effects and therefore could be taken into account to refine user profiles. Further research is needed on these moderators and mediators to use them as criteria for construction of user profiles. Although such research can be complex and costly, it is relevant and useful.

Need of Studies in Implementation and Adoption Stages

The implementation of mHealth solutions (beyond pilots) in the market and their user adoption stages introduce new contextual and technological factors, such as low technical performance, lack of interoperability with existing systems, or misfit with existing clinical practices. Pilot studies often benefit from specific resources—both financial and human—and from high motivation of the patients and health care professionals involved. These factors may be missing in the latter stages of implementation and adoption, influencing the outcome achievement or introducing risks to patient safety. Although these challenges cannot be experimentally controlled, they may be assessed cautiously in rigorous qualitative and statistical studies. In addition, the moderators and mediators of mHealth interventions, such as engagement levels [10], should be investigated more systematically.

We have to mention that the EQUATOR network published guidelines for the reporting of mobile phone–based health interventions [11]. Formed by the World Health Organization's panel of experts, these guidelines are useful to improve the transparency and harmonization of the reporting of mHealth, enabling comparisons and meta-reviews. These criteria include usability/content testing and user feedback. Nevertheless, a new step must be taken toward formulating guidelines on study designs and methods of the assessment of user feedback on mHealth interventions.

Moving Forward: Formal Guidelines for Study Designs on Real-World Usage

The evaluation of moderators and mediators in pilots should be followed by larger surveys and follow-ups during the whole life cycle of the mHealth technology [5]. The protocols of these studies should be inspired by the rigor of protocols of clinical investigations while considering the relevant factors specific to mHealth. Open trials, observational studies, and case reports should be conducted to measure mediators beyond a specific clinical setting (the variables and low sample size of which could introduce serious bias). In addition, anecdotal reports and qualitative studies should use common frameworks, which will facilitate systematic reviews and afford transferability. The knowledge generated can thus inform policy decisions.

Moreover, anecdotal reports of suspected adverse reactions related to the use of mHealth (along with reports of health IT–related adverse events) should be encouraged and published in medical journals, as mHealth can induce specific errors related to the use of technology. These issues have been emphasized as crucial for more than 20 years, during which

numerous studies have shown that bad informatics can have fatal consequences [5]. If the new European regulations on medical devices (which include mHealth solutions) require real-life, postmarket, clinical, and risk assessment follow-ups of these devices, new methods and frameworks should be elaborated with scientific rigor and inspired by different academic disciplines, such as engineering and social sciences. Building on the guidelines for research presentations and knowledge from medical informatics and risk engineering [12], it is time to make rigorous evaluations and formalize guidelines for research presentations, enabling evidence-based mHealth interventions.

Conflicts of Interest

None declared.

References

1. US Food and Drug Administration. Policy for Device Software Functions and Mobile Medical Applications: Guidance for Industry and Food and Drug Administration staff. 2019. URL: https://www.fda.gov/media/80958/download [accessed 2020-08-15]
2. Free C, Phillips G, Watson L, Galli L, Felix L, Edwards P, et al. The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. PLoS Med 2013;10(1):e1001363 [FREE Full text] [doi: 10.1371/journal.pmed.1001363] [Medline: 23458994]
3. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry 2002 Oct;59(10):877-883. [doi: 10.1001/archpsyc.59.10.877] [Medline: 12365874]
4. Lewis TL, Wyatt JC. mHealth and mobile medical Apps: a framework to assess risk and promote safer use. J Med Internet Res 2014 Sep 15;16(9):e210 [FREE Full text] [doi: 10.2196/jmir.3133] [Medline: 25223398]
5. Ammenwerth E, Shaw NT. Bad health informatics can kill--is evaluation the answer? Methods Inf Med 2005;44(1):1-3. [Medline: 15778787]
6. EQUATOR Network. URL: https://www.equator-network.org/ [accessed 2020-08-15]
7. Zapata BC, Fernández-Alemán JL, Idri A, Toval A. Empirical studies on usability of mHealth apps: a systematic literature review. J Med Syst 2015 Feb;39(2):1. [doi: 10.1007/s10916-014-0182-2] [Medline: 25600193]
8. Keutzer L, Simonsson US. Medical Device Apps: An Introduction to Regulatory Affairs for Developers. JMIR Mhealth Uhealth 2020 Jun 26;8(6):e17567 [FREE Full text] [doi: 10.2196/17567] [Medline: 32589154]
9. Borsci S, Londei A, Federici S. The Bootstrap Discovery Behaviour (BDB): a new outlook on usability evaluation. Cogn Process 2011 Feb;12(1):23-31. [doi: 10.1007/s10339-010-0376-6] [Medline: 21046191]
10. Yang Q, Van Stee SK. The Comparative Effectiveness of Mobile Phone Interventions in Improving Health Outcomes: Meta-Analytic Review. JMIR Mhealth Uhealth 2019 Apr 03;7(4):e11244 [FREE Full text] [doi: 10.2196/11244] [Medline: 30942695]
11. Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, WHO mHealth Technical Evidence Review Group. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016 Mar 17;352:i1174. [doi: 10.1136/bmj.i1174] [Medline: 26988021]
12. Jones R, Mateer J. Indirect risk related failures of Medical Information Systems. 2020 Aug 15 Presented at: 14th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE; 2019; Xi'an, China p. 994-999. [doi: 10.1109/iciea.2019.8834057]

Abbreviations

EQUATOR: Enhancing the QUAlity and Transparency Of health Research
FDA: Food and Drug Administration
IT: informational technology
JAMA: Journal of the American Medical Association
mHealth: mobile health
RCT: randomized clinical trial

Edited by C Lovis; submitted 27.06.20; peer-reviewed by R Marcilly, Q Yang; comments to author 03.08.20; revised version received 05.09.20; accepted 12.09.20; published 18.11.20.

Please cite as:
Ologeanu-Taddei R
Assessment of mHealth Interventions: Need for New Studies, Methods, and Guidelines for Study Designs
JMIR Med Inform 2020;8(11):e21874
URL: http://medinform.jmir.org/2020/11/e21874/
doi:10.2196/21874
PMID:33206060

Original Paper

Natural Language Processing for Surveillance of Cervical and Anal Cancer and Precancer: Algorithm Development and Split-Validation Study

Carlos R Oliveira[1], MD, PhD; Patrick Niccolai[1]; Anette Michelle Ortiz[1], BSc; Sangini S Sheth[2], MD, MPH; Eugene D Shapiro[1,3], MD; Linda M Niccolai[3], PhD; Cynthia A Brandt[4,5], MD, MPH

[1]Department of Pediatrics, Yale University School of Medicine, New Haven, CT, United States

[2]Department of Obstetrics, Gynecology, and Reproductive Sciences, Yale University School of Medicine, New Haven, CT, United States

[3]Departments of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, United States

[4]Departments of Emergency Medicine, Biostatistics, and Health Informatics, Yale Schools of Medicine and Public Health, New Haven, CT, United States

[5]Veteran Affairs Connecticut Healthcare System, West Haven, CT, United States

Corresponding Author:
Carlos R Oliveira, MD, PhD
Department of Pediatrics
Yale University School of Medicine
P.O. Box 208000
New Haven, CT, 06520
United States
Phone: 1 203 785 5474
Email: carlos.oliveira@yale.edu

Abstract

Background: Accurate identification of new diagnoses of human papillomavirus–associated cancers and precancers is an important step toward the development of strategies that optimize the use of human papillomavirus vaccines. The diagnosis of human papillomavirus cancers hinges on a histopathologic report, which is typically stored in electronic medical records as free-form, or unstructured, narrative text. Previous efforts to perform surveillance for human papillomavirus cancers have relied on the manual review of pathology reports to extract diagnostic information, a process that is both labor- and resource-intensive. Natural language processing can be used to automate the structuring and extraction of clinical data from unstructured narrative text in medical records and may provide a practical and effective method for identifying patients with vaccine-preventable human papillomavirus disease for surveillance and research.

Objective: This study's objective was to develop and assess the accuracy of a natural language processing algorithm for the identification of individuals with cancer or precancer of the cervix and anus.

Methods: A pipeline-based natural language processing algorithm was developed, which incorporated machine learning and rule-based methods to extract diagnostic elements from the narrative pathology reports. To test the algorithm's classification accuracy, we used a split-validation study design. Full-length cervical and anal pathology reports were randomly selected from 4 clinical pathology laboratories. Two study team members, blinded to the classifications produced by the natural language processing algorithm, manually and independently reviewed all reports and classified them at the document level according to 2 domains (diagnosis and human papillomavirus testing results). Using the manual review as the gold standard, the algorithm's performance was evaluated using standard measurements of accuracy, recall, precision, and F-measure.

Results: The natural language processing algorithm's performance was validated on 949 pathology reports. The algorithm demonstrated accurate identification of abnormal cytology, histology, and positive human papillomavirus tests with accuracies greater than 0.91. Precision was lowest for anal histology reports (0.87, 95% CI 0.59-0.98) and highest for cervical cytology (0.98, 95% CI 0.95-0.99). The natural language processing algorithm missed 2 out of the 15 abnormal anal histology reports, which led to a relatively low recall (0.68, 95% CI 0.43-0.87).

Conclusions: This study outlines the development and validation of a freely available and easily implementable natural language processing algorithm that can automate the extraction and classification of clinical data from cervical and anal cytology and histology.

(JMIR Med Inform 2020;8(11):e20826) doi:10.2196/20826

KEYWORDS

natural language processing; automated data extraction; human papillomavirus; surveillance; pathology reporting; cervical cancer; anal cancer; precancer; cancer; HPV; accuracy

Introduction

Precision public health is a rapidly evolving field that focuses on promoting the health of a population through the application of technology [1]. A key priority in precision public health is the development of new informatics approaches to optimize the use of vaccines for the prevention of disease. Some of the more successful vaccine informatics applications postlicensure include using text-mining techniques to automate the tracking of adverse immunization outcomes and the use of emergency department notes as an early warning sign for outbreaks of vaccine-preventable diseases. Automation of biosurveillance and timely identification of infectious diseases is of particular importance to public health, as it allows for better planning and distribution of limited resources [2-4].

Persistent infection with human papillomavirus (HPV) can result in precancerous anogenital lesions as well as invasive cancer. In the United States, approximately 25,000 cases of anogenital cancers are diagnosed every year, with cervical and anal cancer being the majority (75%) of these [5]. Over 90% of these cases are attributable to infection with HPV types that are preventable by the use of recommended HPV vaccines [5-7]. Although HPV vaccines have high proven efficacy, the way we use these vaccines to prevent HPV cancers is still in need of improvement [8]. Accurate identification and tracking of new cases of HPV cancers is an important step toward the development of strategies that optimize the use of HPV vaccines.

Surveillance for HPV-associated outcomes is critical for monitoring the progress of immunization programs and identifying targets for improvement. Surveillance for HPV cancers, however, has been a formidable challenge. Most of the clinical data needed to diagnose a patient with an HPV-related cancer, or precancer, are stored in pathology reports. Normally, pathology reports are stored in a narrative format and contain several lines of text that can include nondiagnostic information, such as medical history or clinical indications for screening [9]. Although a manual review of these free-text pathology reports is the most accurate case-finding method, it is a laborious process that can become too impractical for large-scale surveillance projects. To facilitate data capture and analysis, considerable efforts have been made to promote processes that encourage pathologists to document their findings in a specific format and using standardized terminology [10]. However, most efforts to incorporate standardized reporting have yet to be consistently implemented by health care providers and institutions [11].

To develop an accurate and scalable surveillance platform for HPV vaccine-preventable cancers, it is critical to first overcome the challenge of narrative data-abstraction. A potential solution to this data-abstraction problem is automation with computational tools, such as natural language processing (NLP). NLP is an increasingly used approach that combines informatics and linguistic techniques to automatically identify and extract key concepts or phrases embedded in a narrative text [12]. Although NLP has been successfully applied for the surveillance of several cancers (eg, colon, hepatic, and bladder cancer), it has been underutilized for the surveillance of HPV cancers and precancers [12-15].

As a first step toward achieving automated surveillance of HPV vaccine-preventable diseases, we developed an NLP algorithm aiming to extract information from cervical and anal pathology reports and classify these reports based on the pathologist's final diagnosis. The objective of this study was to assess the accuracy of our NLP algorithm for the identification of individuals with cancer or precancer of the cervix and anus.

Methods

Study Design and Setting

This study used data generated from the HPV Vaccine Effectiveness Project, a large-scale population-based study aiming to determine the effectiveness of the HPV vaccine [16]. In support of this ongoing project, an NLP algorithm was developed to convert narrative pathology reports into structured data that can be queried to identify individuals who had HPV-related abnormalities in their cervical or anal pathology report. To build and evaluate this NLP algorithm, a split-validation method was used, wherein 2 sets of full-length cervical and anal pathology reports were randomly selected from 4 different clinical pathology laboratories within the Yale–New Haven Health System participating in the HPV Vaccine Effectiveness Project. The first set of reports was used to build the algorithm (ie, the training set, n=100), and the second set was used for testing the accuracy of the algorithm (ie, the validation set, n=1000). Pathology reports were extracted between January 1, 2010 and December 31, 2018 and deidentified for both the development and testing phases of this study.

NLP Algorithm Development

We developed a pipeline-based NLP algorithm that incorporated both machine learning and rule-based methods to extract and classify diagnostic elements (histopathology, cytopathology, and HPV test results) from narrative pathology reports. Various software platforms have been developed to automatically annotate and process clinical notes based on the Unstructured Information Management Architecture framework [17-19]. Our pipeline was built using CLAMP (Clinical Language Annotation, Modeling, and Processing) software, because it is open-source, modular, free-to-use, and specifically designed to process and analyze clinical text [20]. Our pipeline combined several existing and well-validated text processing components [21-27] and built on these components with newly developed HPV-specific ontologies and postprocessing features.

NLP Data Extraction

The first steps of our pipeline involved using CLAMP's existing algorithms to preprocess each report and apply a series of if-then rules to parse and enumerate each sentence and word within the full-length report (ie, a sentence detector and word tokenizer, respectively) [24]. Next, we used a supervised machine learning approach to assign each enumerated token (ie, each word or set of words) a tag based on its part of speech (eg, verb, noun, etc) [28]. A more in-depth description of the pipeline's individual preprocessing components can be found in Multimedia Appendix 1. We then implemented an existing named entity recognizer program to identify key concepts within the narrative text [29]. This named entity recognizer program utilizes a dictionary-based approach to match concepts in pathology reports to terms in a dictionary derived from the Unified Medical Language System Metathesaurus [27]. To more robustly account for variations in HPV-related concepts, we also constructed an HPV-cancer dictionary and incorporated it into the algorithm. This custom HPV-cancer dictionary leveraged over a decade of experience and expertise in HPV-cancer surveillance through collaboration with seasoned epidemiologists from HPV Vaccine Impact Monitoring Project Across Connecticut, a collaborative project between the Connecticut Emerging Infections Program at Yale School of Public Health; the Connecticut Department of Public Health; and the Centers for Disease Control and Prevention [30]. We have contributed our HPV dictionary (ie, ontology) to the National Center for Biomedical Ontology BioPortal platform [31], where it is openly available for other users to develop further.

NLP Data Classification

After implementing the dictionary-based named entity recognizer, we applied newly developed heuristic rules to analyze and relabel each concept based on their context in the report. For example, a series of if-then rules were employed to identify different sections of the report (eg, clinical history, molecular diagnosis, primary diagnosis, etc) and determine when an HPV-related diagnosis was being stated in the report as a historical piece of information and when it was being stated in the context of the current specimen. Further details and examples of the key if-then rules are shown in Multimedia Appendix 1.

We also implemented an extensively validated rule-based negation algorithm [23] to allow us to differentiate when a recognized concept was being negated or stated with uncertainty based on the words that preceded or followed the identified concept (eg, "negative for abnormalities" or "abnormalities were not found"). Once all entities were named, coded, and contextualized, the algorithm generated a structured output (matrix) that was suitable for further processing. For the last step of the algorithm, the structured output was used to summarize and classify each report, at the document level, in 2 key domains: final diagnosis (using the Bethesda Classification system) and results of HPV tests (if performed). To enable the reproducibility of this study, our pipeline was freely available for research through CLAMP [32] and is archived [33]. To facilitate its application, we also provide a step-by-step video demonstration of this pipeline [33].

Classification Validation

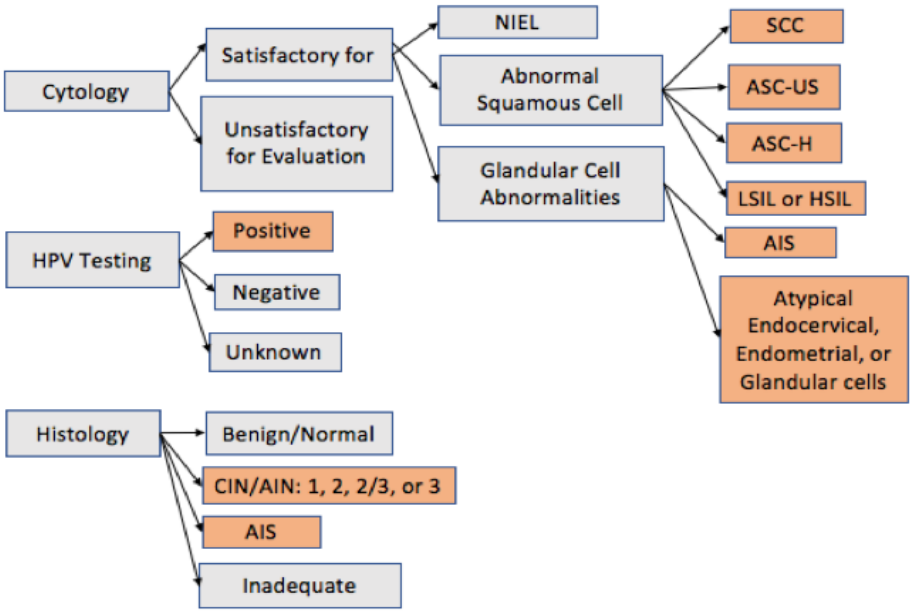
To test the algorithm's classification accuracy, 2 study team members, blinded to the classifications produced by the NLP algorithm, manually and independently reviewed all pathology reports in the validation set and classified them at the document level according to the same 2 domains (diagnosis and HPV testing results). Disagreement among the 2 manual-review adjudicators was resolved by discussion with a third investigator.

For the primary analysis, we tested this algorithm's accuracy for the identification of HPV-related pathology. The primary outcome—abnormal pathology—was grouped as a dichotomous variable and defined, for cytology reports, as a final diagnosis of atypical squamous cells or greater, and for histology reports, as intraepithelial neoplasia grades 2 or greater. A summary of the classification process for the primary outcome is shown in Figure 1.

Statistical Analysis

The algorithm's performance was evaluated using the manual review classifications as the standard. Accuracy, precision, recall, and F-measure were calculated as follows: $accuracy = (true\ positives + true\ negatives) / (true\ positives + true\ negatives + false\ positives + false\ negatives)$; $precision = true\ positives / (true\ positives + false\ positives)$; $recall = true\ positives / (true\ positives + false\ negatives)$; $F\text{-}measure = 2 \times (precision \times recall) / (precision + recall)$. Statistical analyses were conducted using Stata statistical software (version 15; StataCorp LLC). This protocol was approved by the institutional review board of Yale University (protocol number 2000024708).

Figure 1. Diagrammatic representation of the classification process for pathology reports (colored indicates abnormal pathology). AIN: anal intraepithelial lesion; AIS: adenocarcinoma in situ; ASC-US: atypical squamous cells of undetermined significance; ASC-H: atypical squamous cells—cannot exclude high-grade squamous intraepithelial lesion; CIN: cervical intraepithelial lesion;HSIL: high-grade squamous intraepithelial lesion; LSIL: low-grade squamous intraepithelial lesion; NIEL: negative for intraepithelial lesion; SCC: squamous cell carcinoma.



Results

Out of 1000 pathology reports originally selected for the validation set, 51 were excluded after manual review because they were (1) reports with misclassified specimens (ie, not anal or cervical tissue), (2) duplicate reports, or (3) incomplete reports. Testing of the NLP algorithm's accuracy was performed on 949 pathology reports (anal cytology n=94; anal histology n=86; cervical cytology n=403; cervical histology n=366). HPV tests were documented on 303 reports (cervical n=265; anal cytology n=38), of which 121 (40%) had positive results for HPV. A summary of the highest-grade diagnosis based on manual review of the 949 pathology reports is shown in Table 1. Most of the biopsies performed revealed either normal or low-grade (362/452, 80%) lesions, and most of the cytologic specimens were negative for intraepithelial lesions (302/497, 61%).

Table 1. Summary of results from the manual review of the validation set.

Test	Cervical (n=769), n (%)	Anal (n=180), n (%)	Total (N=949), n
Cytology	403 (81.1)	94 (18.9)	497
Negative for intraepithelial lesion	255 (84.4)	47 (15.6)	302
Atypical squamous cells of undetermined significance	44 (68.8)	20 (31.3)	64
Atypical squamous cells—cannot exclude high-grade squamous intraepithelial lesion	57 (98.3)	1 (1.7)	58
Low-grade squamous intraepithelial lesion	16 (84.2)	3 (15.8)	19
Glandular abnormality	14 (82.4)	3 (17.6)	17
Unsatisfactory specimen	17 (45.9)	20 (54.1)	37
HPV[a] test performed	206 (85.8)	34 (14.2)	240
Positive	91 (84.3)	17 (15.7)	108
Histology	366 (81.0)	86 (19.0)	452
Benign	153 (77.3)	45 (22.7)	198
Squamous intraepithelial lesion grade 1	138 (84.1)	26 (15.9)	164
Squamous intraepithelial lesion grade 2+	75 (83.3)	15 (16.7)	90

[a]HPV: human papillomavirus.

For the primary analysis, the NLP algorithm accurately identified abnormal cytology, histology, and positive HPV tests with accuracies ≥0.91 in all specimens (Table 2). Precision was lowest for anal histology reports (0.87, 95% CI 0.59-0.98) and highest for cervical cytology (0.98, 95% CI 0.95-0.99). The NLP algorithm missed 2 out of the 15 abnormal anal histology

reports, which led to relatively low recall (0.68, 95% CI 0.43-0.87).

Table 2. Performance of NLP algorithm on the validation set, N = 949.

Variable	Precision (95% CI)	Recall (95% CI)	F-measure (95% CI)	Accuracy (95% CI)
Abnormal cytology[a]				
Cervical	0.98 (0.95-0.99)	1.00 (0.97-1.00)	0.99 (0.98-1.00)	0.99 (0.98-1.00)
Anal	0.93 (0.76-0.99)	1.00 (0.86-1.00)	0.96 (0.91-1.00)	0.98 (0.93-0.99)
HPV[b] **testing**				
Positive	0.95 (0.89-0.98)	1.00 (0.97-1.00)	0.97 (0.95-0.99)	0.99 (0.98-1.00)
Abnormal histology				
CIN[c] grade 2+	0.89 (0.80-0.95)	0.93 (0.85-0.98)	0.91 (0.86-0.96)	0.96 (0.94-0.98)
AIN[d] grade 2+	0.87 (0.59-0.98)	0.68 (0.43-0.87)	0.76 (0.61-0.92)	0.91 (0.82-0.96)
Average performance[e]				
Abnormal test	0.94 (0.91-0.97)	0.96 (0.92-0.98)	0.94 (0.93-0.97)	0.97 (0.96-0.98)

[a]Abnormalities include atypical squamous cells of undetermined significance, atypical squamous cells—cannot exclude high-grade squamous intraepithelial lesion, low-grade squamous intraepithelial lesion, and glandular cell abnormalities.

[b]HPV: human papillomavirus.

[c]CIN: cervical intraepithelial lesion.

[d]AIN: anal intraepithelial lesion.

[e]Includes results from both cytology and histology.

Discussion

In this paper, we described the development and validation of an NLP instrument that can be used for both data extraction and classification of cytology and histology reports of the cervix and anus. Based on these initial data, our NLP algorithm can classify whether a cytology or histology specimen was abnormal and whether any HPV tests resulted positive, with an accuracy 91%. At the document level, this algorithm had an average recall (also known as sensitivity) of 96% and precision (also known as positive predictive value) of 94%. This demonstration of accuracy is an important first step toward the development of a tool that can facilitate the automation of surveillance for HPV vaccine-preventable cancers and precancers.

There is an increasing body of evidence showing the merits of an NLP system over manual review for data extraction and document classification for disease surveillance [34,35]. A key contribution of this study is the integration and application of well-validated NLP methodologies to solve a real-world public health problem. Most individual components included in our NLP pipeline have been previously validated. Using a commonly used corpus (SemEval-2014), Soysal et al [20] demonstrated that CLAMP's named entity recognizer algorithm had superior precision to those of other commonly used platforms (CLAMP: 0.77; MetaMa*P*: 0.55; cTAKES: 0.46). In the same study [20], the performance accuracy of other key components (tokenizer, sentence boundary detector, part-of-speech tagger, and section detector) were evaluated using the MiPACQ clinical corpus and were also found to have a high accuracy (>92%). In our study, we did not aim to develop novel NLP strategies or components. However, one of the key strengths in our approach is that we were able to leverage the experience of HPV surveillance experts to assemble an extensive list of HPV-related terms to optimize named entity recognition.

This study has several other notable strengths. First, this study is among the first to evaluate the accuracy of an NLP algorithm to identify cases of HPV-related precancers. Although precancerous diagnoses are routinely made, these data are not systematically collected by most surveillance systems. These diagnoses, however, have public health significance as they can be used to monitor the impact of HPV vaccines. Our NLP algorithm provides an efficient way to use existing resources to measure the extent to which HPV vaccines reduce the burden of disease at the population level and identify areas to strengthen immunization programs. Automating the identification of precancers may also have clinical applications. For example, following an abnormal cytology result, a patient is usually kept under close surveillance for months. After an abnormal cytology screen, the appropriate management can vary from more frequent follow-up tests to immediate treatment with surgical excision. Automation of the detection of precancerous abnormalities in cytology or histology can be incorporated into clinical decision support tools to ensure patients are appropriately linked to care and are receiving timely follow-up.

An additional strength of this study is in the application of our NLP algorithm to accurately detect cases of anal cancer and precancer. To our knowledge, we are the first to provide a tool specifically designed for this purpose. Efforts to monitor the impact of HPV vaccination on oncogenic outcomes have focused mainly on cervical cancer and women. With the increased recognition that HPV also causes cancer in men and the increasing rate of these cancers in the young adult population [25,26], it is important to determine if the HPV vaccine's deployment can be optimized to reduce the burden of disease

in both sexes. A surveillance system with these outcomes may be especially valuable to investigators and public health officials in assessing the impact of various immunization strategies in both males and females.

Additional improvements can optimize the performance of this algorithm for implementation in routine public health surveillance or clinical practice. For example, we only used reports from a single health care system (Yale New Haven Health), which likely limited the variability found in both the structure and language in the pathology reports. Thus, future work is needed to validate this tool's portability to other health care systems where pathology practices may differ. An additional area of improvement is in the preprocessing. After initial manual review of pathology reports, we had to exclude several reports that were incomplete or were misclassified in the electronic medical record. To be useful as a real-time surveillance tool, future iterations of this NLP algorithm will need to address the potential for misclassification at the onset. An additional limitation of this tool is that it was developed as a means to identify cases of cancer and precancer at the document level and not at the patient level. As many individuals have more than one pathology report in their record, to be useful as an automated surveillance method, more postprocessing will be needed to deal with duplicates or disparate findings at the patient level.

In this study, we detail the development of a freely available and easily implementable NLP algorithm that can automate the extraction of clinical data from cervical and anal cytology and histology reports. We show that with this algorithm, it is possible to accurately detect patients with HPV-related abnormalities at these anatomical sites. These data provide preliminary support for the use of our NLP instrument for the surveillance of HPV cancer and precancer of the cervix and anus.

Acknowledgments

We would like to acknowledge the team of investigators and epidemiologists at HPV Vaccine Impact Monitoring Project Across Connecticut: Kyle Higgins, Monica Brackney, and James Meek.

This work was supported in part by National Institutes of Health grant number R01AI123204 (PN) from the National Institute of Allergy and Infectious Diseases and grant numbers KL2TR001862 (CRO) and UL1TR000142 (EDS) from the National Center for Advancing Translational Science. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of National Institutes of Health. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all of the data in this study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

LMN reports previous work as a scientific advisor for Merck. SSS has previously provided consulting services to Merck and received a research grant from Merck. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Supplementary methods.
[DOCX File , 894 KB - medinform_v8i11e20826_app1.docx]

References

1. Bayer R, Galea S. Public health in the precision-medicine era. N Engl J Med 2015 Aug 06;373(6):499-501. [doi: 10.1056/NEJMp1506241] [Medline: 26244305]
2. Yu W, Zheng C, Xie F, Chen W, Mercado C, Sy LS, et al. The use of natural language processing to identify vaccine-related anaphylaxis at five health care systems in the Vaccine Safety Datalink. Pharmacoepidemiol Drug Saf 2020 Feb;29(2):182-188. [doi: 10.1002/pds.4919] [Medline: 31797475]
3. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Ann Intern Med 2012 Jan 03;156(1 Pt 1):11-18. [doi: 10.7326/0003-4819-156-1-201201030-00003] [Medline: 22213490]
4. Ye Y, Tsui FR, Wagner M, Espino JU, Li Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J Am Med Inform Assoc 2014;21(5):815-823 [FREE Full text] [doi: 10.1136/amiajnl-2013-001934] [Medline: 24406261]
5. Centers for Disease Control and Prevention. Cancers Associated with Human Papillomavirus, United States—2012–2016. USCS Data Brief.: US Department of Health and Human Services; 2019. URL: www.cdc.gov/cancer/uscs/about/data-briefs/no10-hpv-assoc-cancers-UnitedStates-2012-2016.htm [accessed 2020-10-27]
6. Petrosky E, Bocchini JA, Hariri S, Chesson H, Curtis CR, Saraiya M, Centers for Disease Control and Prevention. Use of 9-valent human papillomavirus (HPV) vaccine: updated HPV vaccination recommendations of the advisory committee on immunization practices. MMWR Morb Mortal Wkly Rep 2015 Mar 27;64(11):300-304 [FREE Full text] [Medline: 25811679]

7. Gargano J, Meites E, Watson M, Unger E, Markowitz L. Chapter 5: human papillomavirus. In: Roush SW, Baldy LM, Kirkconnell Hall MA, editors. Manual for the Surveillance of Vaccine-Preventable Diseases. Atlanta, GA: Centers for Disease Control and Prevention Department of Health and Human Services; Apr 28, 2020.
8. Sivaram S, Sanchez MA, Rimer BK, Samet JM, Glasgow RE. Implementation science in cancer prevention and control: a framework for research and programs in low- and middle-income countries. Cancer Epidemiol Biomarkers Prev 2014 Nov;23(11):2273-2284. [doi: 10.1158/1055-9965.EPI-14-0472] [Medline: 25178984]
9. Crothers BA, Tench WD, Schwartz MR, Bentz JS, Moriarty AT, Clayton AC, et al. Guidelines for the reporting of nongynecologic cytopathology specimens. Arch Pathol Lab Med 2009 Nov;133(11):1743-1756. [doi: 10.1043/1543-2165-133.11.1743] [Medline: 19886707]
10. Renshaw AA, Mena-Allauca M, Gould EW, Sirintrapun SJ. Synoptic reporting: evidence-based review and future directions. JCO Clin Cancer Inform 2018 Dec;2:1-9. [doi: 10.1200/CCI.17.00088] [Medline: 30652566]
11. Bodenreider O, Cornet R, Vreeman DJ. Recent developments in clinical terminologies - SNOMED CT, LOINC, and RxNorm. Yearb Med Inform 2018 Aug;27(1):129-139 [FREE Full text] [doi: 10.1055/s-0038-1667077] [Medline: 30157516]
12. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clin Gastroenterol Hepatol 2013 Jun;11(6):689-694 [FREE Full text] [doi: 10.1016/j.cgh.2012.11.035] [Medline: 23313839]
13. Wagholikar KB, MacLaughlin KL, Henry MR, Greenes RA, Hankey RA, Liu H, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc 2012;19(5):833-839. [doi: 10.1136/amiajnl-2012-000820] [Medline: 22542812]
14. Schroeck FR, Patterson OV, Alba PR, Pattison EA, Seigne JD, DuVall SL, et al. Development of a natural language processing engine to generate bladder cancer pathology data for health services research. Urology 2017 Dec;110:84-91 [FREE Full text] [doi: 10.1016/j.urology.2017.07.056] [Medline: 28916254]
15. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural language processing technologies in radiology research and clinical applications. Radiographics 2016;36(1):176-191 [FREE Full text] [doi: 10.1148/rg.2016150080] [Medline: 26761536]
16. Oliveira CR. Estimating the Effectiveness of Human Papillomavirus Vaccine: A Case-Control Study with Bayesian Model Averaging. In: Yale University. Ann Arbor: ProQuest Dissertations & Theses Global; 2019:126.
17. Bates J, Fodeh SJ, Brandt CA, Womack JA. Classification of radiology reports for falls in an HIV study cohort. J Am Med Inform Assoc 2016 Apr;23(e1):e113-e117 [FREE Full text] [doi: 10.1093/jamia/ocv155] [Medline: 26567329]
18. Garla V, Lo RV, Dorey-Stein Z, Kidwai F, Scotch M, Womack J, et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 2011;18(5):614-620. [doi: 10.1136/amiajnl-2011-000093] [Medline: 21622934]
19. Womack JA, Murphy TE, Rentsch CT, Tate JP, Bathulapalli H, Smith AC, et al. Polypharmacy, hazardous alcohol and illicit substance use, and serious falls among PLWH and uninfected comparators. J Acquir Immune Defic Syndr 2019 Nov 01;82(3):305-313. [doi: 10.1097/QAI.0000000000002130] [Medline: 31339866]
20. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2018 Mar 01;25(3):331-336. [doi: 10.1093/jamia/ocx132] [Medline: 29186491]
21. Le DV, Montgomery J, Kirkby KC, Scanlan J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. J Biomed Inform 2018 Oct;86:49-58 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.007] [Medline: 30118855]
22. Redman JS, Natarajan Y, Hou JK, Wang J, Hanif M, Feng H, et al. Accurate identification of fatty liver disease in data warehouse utilizing natural language processing. Dig Dis Sci 2017 Oct;62(10):2713-2718. [doi: 10.1007/s10620-017-4721-9] [Medline: 28861720]
23. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001 Oct;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
24. Doan S, Bastarache L, Klimkowski S, Denny JC, Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. J Am Med Inform Assoc 2010 Oct;17(5):528-531. [doi: 10.1136/jamia.2010.003855] [Medline: 20819857]
25. Brotherton JML, Giuliano AR, Markowitz LE, Dunne EF, Ogilvie GS. Monitoring the impact of HPV vaccine in males-considerations and challenges. Papillomavirus Res 2016 Dec;2:106-111 [FREE Full text] [doi: 10.1016/j.pvr.2016.05.001] [Medline: 29074169]
26. Palefsky JM. Human papillomavirus-related disease in men: not just a women's issue. J Adolesc Health 2010 Apr;46(4 Suppl):S12-S19 [FREE Full text] [doi: 10.1016/j.jadohealth.2010.01.010] [Medline: 20307839]
27. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Supp 1):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
28. Apache OpenNLP Development Community. Apache OpenNLP.: The Apache Software Foundation; 2020. URL: http://opennlp.apache.org/index.html [accessed 2020-10-27]

29. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
30. Hariri S, Markowitz LE, Bennett NM, Niccolai LM, Schafer S, Bloch K, Hpv-Impact Working Group. Monitoring effect of human papillomavirus vaccines in US population, emerging infections program, 2008-2012. Emerg Infect Dis 2015 Sep;21(9):1557-1561 [FREE Full text] [doi: 10.3201/eid2109.141841] [Medline: 26291379]
31. National Center for Biomedical Ontology. Bioportal.: The Board of Trustees of Leland Stanford Junior University; 2020. URL: http://bioportal.bioontology.org/ontologies/HPV [accessed 2020-10-27]
32. Hao D. Clinical Language Annotation, Modeling, and Processing Toolkit. Houston, TX: Center for Computational Biomedicine; 2020. URL: http://clamp.uth.edu [accessed 2020-10-27]
33. Niccolai P, Oliveira CR. HPV Pathology CLAMP Pipeline.: GitHub; 2020. URL: https://github.com/PatrickNiccolai1/HPV_Pathology_Clamp_Pipeline [accessed 2020-10-27]
34. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017 Dec;73:14-29 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
35. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. J Am Med Inform Assoc 2016 Nov;23(6):1077-1084. [doi: 10.1093/jamia/ocw006] [Medline: 27026618]

Abbreviations

CLAMP: Clinical Language Annotation, Modeling, and Processing
HPV: human papillomavirus
NLP: natural language processing

Edited by G Eysenbach; submitted 29.05.20; peer-reviewed by S Noah, S Doan, Y Motoki; comments to author 19.06.20; revised version received 18.09.20; accepted 04.10.20; published 03.11.20.

Please cite as:
Oliveira CR, Niccolai P, Ortiz AM, Sheth SS, Shapiro ED, Niccolai LM, Brandt CA
Natural Language Processing for Surveillance of Cervical and Anal Cancer and Precancer: Algorithm Development and Split-Validation Study
JMIR Med Inform 2020;8(11):e20826
URL: https://medinform.jmir.org/2020/11/e20826
doi:10.2196/20826
PMID:32469840

Original Paper

Characterizing Chronic Pain Episodes in Clinical Text at Two Health Care Systems: Comprehensive Annotation and Corpus Analysis

Luke A Carlson[1*], BA; Molly M Jeffery[2*], PhD; Sunyang Fu[1], MHI; Huan He[1], PhD; Rozalina G McCoy[3], MD, MSc; Yanshan Wang[1], PhD; William Michael Hooten[4], MD; Jennifer St Sauver[5], PhD; Hongfang Liu[1], PhD; Jungwei Fan[1], PhD

[1]Division of Digital Health Sciences, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

[2]Division of Health Care Policy and Research, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

[3]Division of Community Internal Medicine, Department of Medicine, Mayo Clinic, Rochester, MN, United States

[4]Division of Pain Medicine, Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, Rochester, MN, United States

[5]Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

[*]these authors contributed equally

Corresponding Author:
Jungwei Fan, PhD
Division of Digital Health Sciences
Department of Health Sciences Research
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 5075381191
Email: fan.jung-wei@mayo.edu

Abstract

Background: Chronic pain affects more than 20% of adults in the United States and is associated with substantial physical, mental, and social burden. Clinical text contains rich information about chronic pain, but no systematic appraisal has been performed to assess the electronic health record (EHR) narratives for these patients. A formal content analysis of the unstructured EHR data can inform clinical practice and research in chronic pain.

Objective: We characterized individual episodes of chronic pain by annotating and analyzing EHR notes for a stratified cohort of adults with known chronic pain.

Methods: We used the Rochester Epidemiology Project infrastructure to screen all residents of Olmsted County, Minnesota, for evidence of chronic pain, between January 1, 2005, and September 30, 2015. Diagnosis codes were used to assemble a cohort of 6586 chronic pain patients; people with cancer were excluded. The records of an age- and sex-stratified random sample of 62 patients from the cohort were annotated using an iteratively developed guideline. The annotated concepts included date, location, severity, causes, effects on quality of life, diagnostic procedures, medications, and other treatment modalities.

Results: A total of 94 chronic pain episodes from 62 distinct patients were identified by reviewing 3272 clinical notes. Documentation was written by clinicians across a wide spectrum of specialties. Most patients (40/62, 65%) had 1 pain episode during the study period. Interannotator agreement ranged from 0.78 to 1.00 across the annotated concepts. Some pain-related concepts (eg, body location) had 100% (94/94) coverage among all the episodes, while others had moderate coverage (eg, effects on quality of life) (55/94, 59%). Back pain and leg pain were the most common types of chronic pain in the annotated cohort. Musculoskeletal issues like arthritis were annotated as the most common causes. Opioids were the most commonly captured medication, while physical and occupational therapies were the most common nonpharmacological treatments.

Conclusions: We systematically annotated chronic pain episodes in clinical text. The rich content analysis results revealed complexity of the chronic pain episodes and of their management, as well as the challenges in extracting pertinent information, even for humans. Despite the pilot study nature of the work, the annotation guideline and corpus should be able to serve as informative references for other institutions with shared interest in chronic pain research using EHRs.

(JMIR Med Inform 2020;8(11):e18659) doi:10.2196/18659

KEYWORDS

chronic pain; guideline development; knowledge representation; corpus annotation; content analysis

Introduction

Significance

Chronic pain (ie, pain persisting for >90 days) can be debilitating to both physical and emotional well-being and has resulted in significant socioeconomic costs [1,2]. In 2016, it was estimated that 20.4% (50.0 million) of US adults had chronic pain, with 8.0% (19.6 million) experiencing high-impact chronic pain that can frequently limit life or work activities [1]. The annual costs of medical treatment, lost productivity, and disability programs have been estimated at US $560 billion in the United States alone [3]. Effective treatment and management of chronic pain is difficult, due to complex and frequently multifactorial etiology, intertwined mental health conditions, and progression of pain from a symptom to a disease in itself [4]. Research is urgently needed to understand chronic pain through real-world data and to inform best practices for patient care.

Electronic health records (EHRs) hold great promise for chronic pain research, offering rich and contextualized practice-generated evidence. EHR data may also facilitate examination of the effectiveness and safety of chronic pain interventions [5], which heretofore has been limited. Unstructured narratives (ie, clinical notes) in EHRs are indispensable to understanding the full context of a patient's experience, and most clinicians prefer the expressiveness of free text in documenting pain [6]. However, there has been no systematic appraisal of EHR narratives surrounding chronic pain. Therefore, as a foundational step toward filling this gap, we annotated and analyzed the clinical notes of patients with chronic pain diagnoses.

Guided by clinical practice and research needs, we annotated information related to body location, severity, causes, social and emotional effects, and interventions associated with chronic pain across a wide range of symptomatology and etiology. We centered around individual episodes of chronic pain, examining notes spanning the period from 6 months before to 2 years after the first chronic pain diagnosis for each patient. A total of 3272 notes were reviewed, and 94 episodes from 62 distinct patients were identified. The results showed that the clinical notes contain valuable information on chronic pain, effectively covering key aspects such as location and cause of pain in more than 90% of the episodes. Moreover, aspects of chronic pain generally not available in structured EHR data, like alternative regimens and the effects on quality of life, also had a sizable presence in the annotated corpus.

Background

Previous work pointed out that chronic pain surveillance can be limited without using information in clinical text [7], and there has been substantial interest in detecting pain phenotypes based on EHR documents. Heintzelman et al [8] used a proprietary, rule-based system to identify mentions of pain and attributes such as location and severity for 33 prostate cancer patients. Two prior studies developed and analyzed clinical corpora on pain and pain management. Dorflinger et al [9] sampled 153 patients with pain scores of 4 or higher in the Veterans Affairs primary care setting and used their progress notes to develop an information extraction schema on the quality of pain management documentation. Their schema identified three major areas—pain assessment, treatment, and reassessment—that covered several indicators such as cause, constancy, and pain sensation. In developing an annotation schema for anesthesia information and postsurgical pain, Yim et al [10] sampled 420 notes from patients that underwent five procedures. Many pain-related attributes aligned with those identified by Dorflinger et al [9] (eg, trigger, location, frequency, and pain character). Together, these studies confirmed that free-text notes contained relevant information for understanding pain and evidence of management approaches. Our study built upon this previous work and focused specifically on chronic pain. In particular, we developed an episode-based framework that allowed us to define each chronic pain entity longitudinally in a meaningful way.

Methods

Cohort Identification

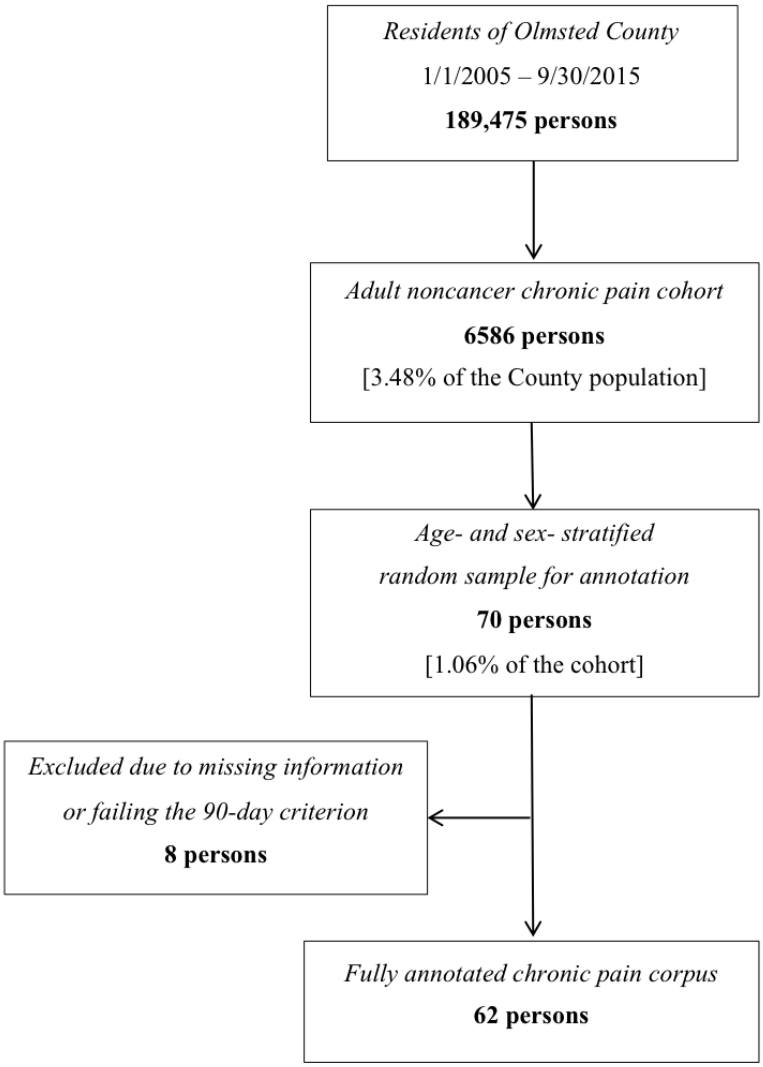
Our study was approved by the Mayo Clinic and the Olmsted Medical Center (OMC) Institutional Review Boards (IRBs). The study population consisted of local adult patients with noncancer chronic pain receiving health care at the Mayo Clinic and/or the OMC. We used the Rochester Epidemiology Project (REP) [11,12] research infrastructure, which covers virtually all residents living in Olmsted County, Minnesota, between January 1, 2005, and September 30, 2015, totaling 189,475 persons. Patients who were coded with any *highly likely* code for chronic pain from the International Classification of Diseases, Ninth Revision (ICD-9), defined by Tian et al [7] (see Table 1), were included. Patients were excluded if they were younger than 19 years of age or if they had any ICD-9 code for cancer between 2003 and 2016; cancer was excluded because cancer-related pain has different treatment patterns. Accordingly, a cohort of 6586 patients out of 189,475 (3.48%) screened persons was established.

Table 1. International Classification of Diseases, Ninth Revision (ICD-9) codes determined by Tian et al [7] as *highly likely* to represent chronic pain.

ICD-9 code	Description
338.2	Chronic pain
338.21	Chronic pain due to trauma
338.22	Chronic post-thoracotomy pain
338.28	Other chronic postoperative pain
338.29	Other chronic pain
338.4	Chronic pain syndrome

To diversify the demographics for more diverse and representative annotation, an age- and sex-stratified sample of 70 patients was randomly selected from the chronic pain cohort. The four age strata were 19-35, 36-50, 51-65, and >65 years of age. Of these 70 selected patients, 8 (11%) were later excluded from annotation due to the absence of any documented pain episode lasting 90 days or longer. An overview of the screening and sampling process is summarized in Figure 1.

Figure 1. Workflow of screening the noncancer chronic pain cohort and sampling for the corpus annotation.



Pain Episode Framework

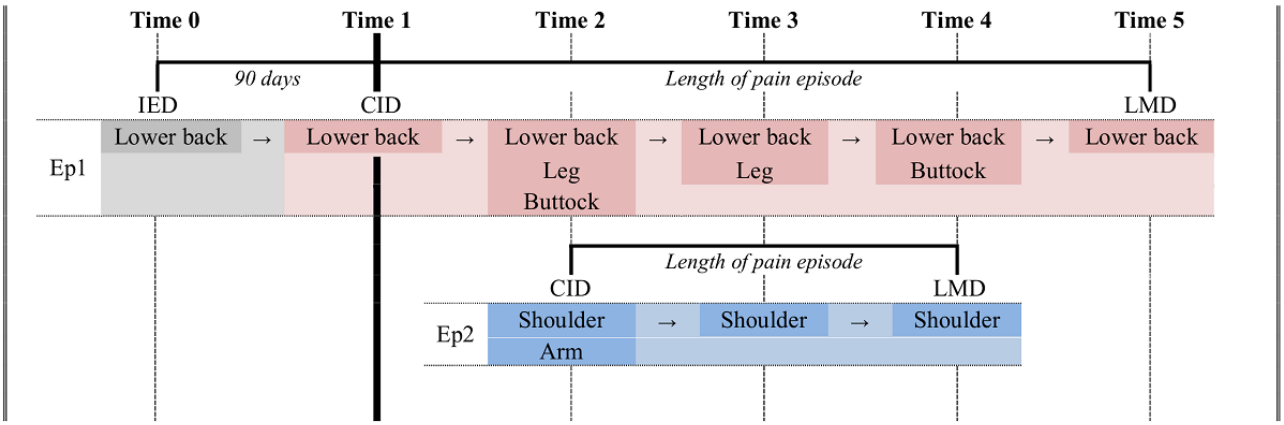
We proposed an episode-centered approach to annotating chronic pain based on input from clinical domain experts. Conceptually, each pain episode involved three points in time:

1. Initial event date: the date on which the pain first presented (eg, as the result of a fall). Since the initiation was not always specified in the EHR, the annotation could be an estimate and was not mandated for defining an episode.

2. Chronic pain index date: the date on which the pain was considered chronic (ie, at least 90 days after the initial event date). If the initial event date was unclear, then the first note implying that the pain had lasted at least 90 days or making explicit use of the word “chronic” was annotated as the chronic pain index date.
3. Last mention date: the date on which the pain was considered to be resolved or the patient was censored. This date could be determined by a note explicitly indicating resolution of the pain, no further mentions of the pain after that note, or a cutoff at 2 years after the index diagnosis if neither of the aforementioned criteria were met.

Operationally, a chronic pain episode was defined by the chronic pain index date and last mention date plus at least one consistent location mentioned over time. If multiple locations were mentioned in the clinical note, annotators used their judgment to determine whether the locations could be physiologically linked. When the locations were all generated from one source (eg, lower back pain and leg pain due to sciatica), all locations were annotated as part of the same episode. As an example, Time 2 in Figure 2 illustrates how five pain locations can be annotated into two separate episodes, where episode 1 had started and evolved in parallel with the later episode 2.

Figure 2. Conceptual representation of a patient’s two chronic pain episodes that unfolded in parallel. Note that the even time intervals are a simplified illustration; the real events have varying intervals. CID: chronic pain index date; Ep1: episode 1; Ep2: episode 2; IED: initial event date; LMD: last mention date.



Corpus Annotation and Analysis

The first chronic pain ICD-9 code from structured data served as the anchor for corpus preparation, however, this diagnosis date might be different from the chronic pain index date determined later by an annotator; all of the patient’s clinical notes 6 months before and 2 years after this anchor diagnosis were then retrieved for annotation. The Multi-document Annotation Environment [13] was the annotation tool of choice because it is lightweight and easy to set up. The primary annotation tasks were (1) screening each patient’s notes to verify that they had at least one pain episode lasting 90 days or longer, (2) determining the pain episode boundary by identification of the chronic pain index date, last mention date, and initial event date, if applicable, and (3) marking mentions of the pain and associated attributes including date, location, severity, cause, effects on life, diagnostic procedure, medication, and other treatment regimens.

During the initial guideline development phase, multiple iterations of revisions were performed on both the guideline and the annotations. Each clinical note was independently annotated by two annotators (LC and MJ), and the interannotator agreement (IAA) was evaluated using the F1 score [14]. After each iteration, disagreements were resolved through discussion with a third reviewer (WH or JS). The common disagreements were logged and analyzed. The annotation guideline stabilized after going through such dual annotation and reconciliation over 604 notes. Following the guideline development, a total of 3272 notes representing 62 patients were reviewed by the two annotators in parallel. Upon completing the corpus annotation, descriptive statistics were computed to summarize the chronic pain episodes and attributes.

Ethics Statement

The research involved secondary use of health records and did not involve a clinical trial. Because the research only used data passively collected as part of clinical care, and did not involve patient contact, the protocol was categorized as minimal risk. The requirement for informed consent was waived by the governing IRBs (Mayo Clinic: 18-006536 and Olmsted Medical Center: 038-OMC-18). We note that while informed consent was not required, Minnesota state law requires that health care providers collect and maintain patient authorization for linking medical record information across health care providers for research. All health care providers participating in the REP maintain research authorization, and we did not include the medical record information for patients who declined research authorization (<5% of potential participants).

Data Availability

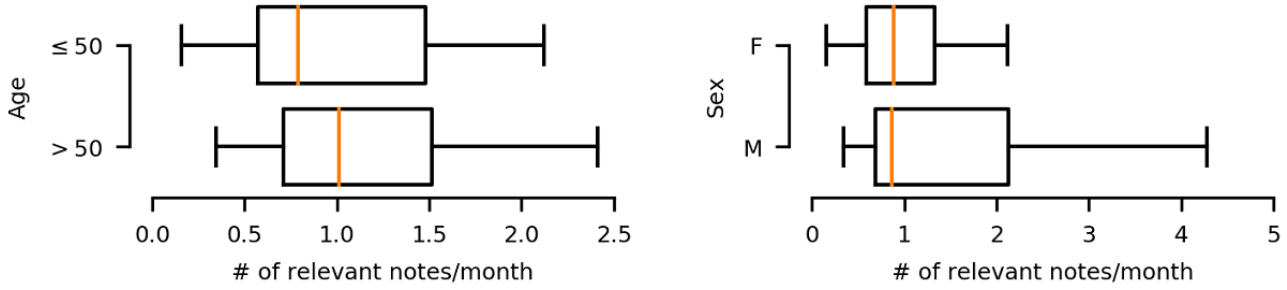
Deidentified annotations of the chronic pain episodes are available for noncommercial research purposes. Interested parties can request access by contacting rstnlp@mayo.edu and are required to sign and remain compliant with a Data Use Agreement under the Mayo Clinic NLP (natural language processing) Program (IRB 20-001137).

Results

Summary of the Annotated Cohort

The final annotated cohort consisted of 62 patients (34/62, 55% female) with balanced representation (approximately 15 patients) from each of the four age strata. To assess the density of relevant documentation, we computed the frequency distributions of the chronic pain–annotated notes per month (see Figure 3). Aligning with intuition, older patients (>50 years) had more clinical notes per month that documented chronic pain. The medians of relevant notes per month for men and women were comparable, but men exhibited a wider variance toward the higher end.

Figure 3. Boxplots for the per-month distributions of the annotated clinical notes, stratified by age (left) and by sex (right). Orange lines indicate medians; boxes are based on interquartile range. F: female; M: male.



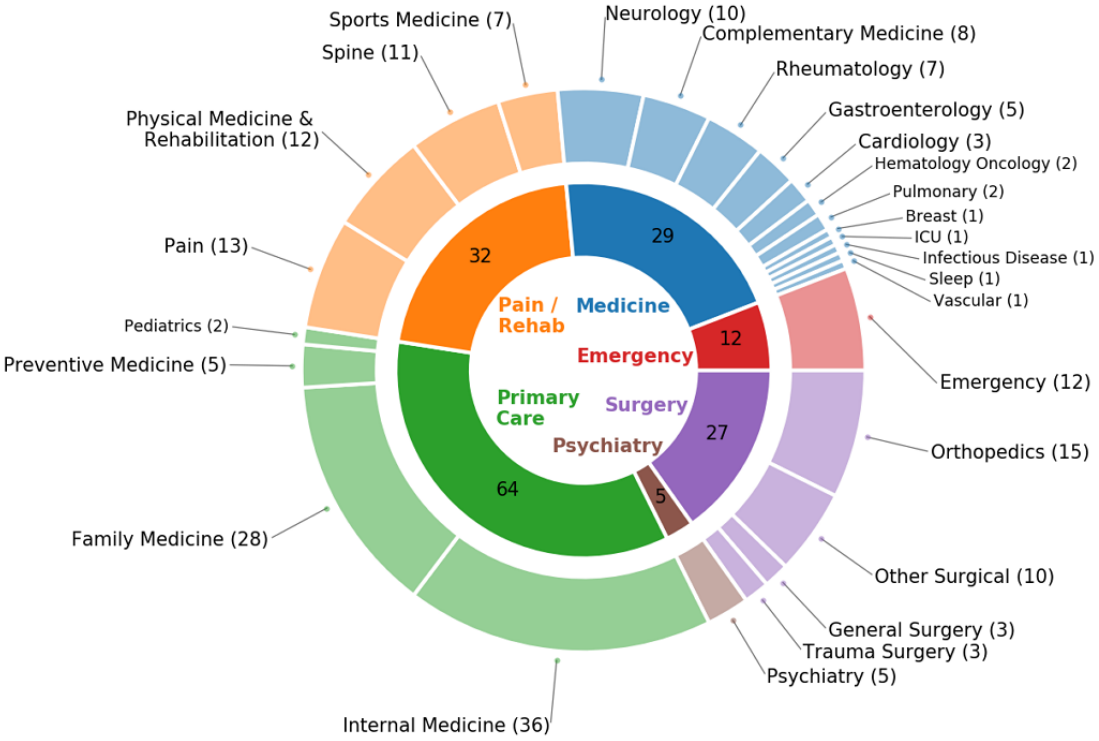
A total of 94 chronic pain episodes were identified from the 62 annotated patients (see Table 2). The median duration of each pain episode was 357 days (min 90 and max 977). To describe which specialties contributed to caring for and documenting the chronic pain patients within this cohort, we computed the specialty-wise episode counts in Figure 4. Primary care departments led the coverage (64/94, 68%), followed by pain and rehabilitation specialties (32/94, 34%) and a variety of medical specialties (29/94, 31%).

Table 2. Frequency distribution of patients per number of annotated chronic pain episodes.

Episodes, n	Patients (N=62), n (%)[a]
1	40 (65)
2	14 (23)
3	5 (8)
4	2 (3)
5	1 (2)

[a]Percentages do not add up to 100% due to rounding.

Figure 4. Counts of the annotated pain episodes by specialty. The inner circle represents the broader specialty category; the outer circle hosts the individual specialties under each category. Each quantity simply indicates how many episodes out of the 94 were cared for by the corresponding specialty or category. Note that an episode could be cared for by more than one specialty, so the counts involve overlaps and the sums may not match across. ICU: intensive care unit; Rehab: rehabilitation.



Summary of the Annotated Corpus

In the guideline development phase, 30 episodes from 604 clinical notes were dually annotated. The IAA rates at the episode and concept levels are provided in Table 3. Note that the concept-level IAA overpenalized because each annotator had the freedom to extract the same information from any legitimate piece of evidence, for the same episode, but those differentially located pieces were counted as mismatches. As our primary focus, the episode-level IAA indicated moderate to high agreement (0.78 to 1.00) and viability of the final annotation guideline (see Multimedia Appendix 1).

Table 3. Interannotator agreement (IAA) rates computed using the F1 score.

Annotation	Episode-level IAA rate	Concept-level IAA rate
Episode	0.89	0.79
Date	1.00	0.94
Location	0.84	0.82
Severity	0.78	0.70
Cause	0.90	0.68
Social and emotional effect	0.78	0.53
Diagnostic procedure	0.78	0.65
Medication	0.82	0.70
Other treatment	0.80	0.65

The individual pain attributes and corresponding examples are summarized in Table 4, along with the percentage of episodes that had the attribute covered. For example, 100% (94/94) of the episodes had *date* annotated, while only 59% (55/94) described the *social and emotional effect* of pain. Example annotations and the number of distinct strings of each annotated attribute are also included in the table. The most frequent subcategories of each annotated attribute are provided in Multimedia Appendix 2 (for *cause*, *social and emotional effect*, *diagnostic procedure*, *medication*, and *other treatment*) and Figure 5 (for *location* and *severity*). Structured export of a mock-up episode is shown in Figure 6, which represents the

commonly included information: patient ID; episode begin and end dates; location, with mapping to SNOMED CT (Systematized Nomenclature Of Medicine–Clinical Terms); severity; medication; document ID; and the character spans of the annotations in text.

Table 4. Chronic pain attributes, examples, distinct strings, and the frequency of episodes that had the attribute annotated.

Attribute	Definition	Examples of annotation	Distinct strings, n	Coverage of episodes (N=94), n (%)
Date	Chronic pain index date, last mention date, or initial event date	"05-23-2009" "January 12, 2010"	174	94 (100)
Location	Body part where the pain occurred	"left knee" "lower back"	240	94 (100)
Severity	Perceived pain intensity	"tolerable" "9/10"	295	79 (84)
Cause	Etiology or contributing factor	"arthritis" "peripheral neuropathy"	390	88 (94)
Social and emotional effect	Consequence to daily life	"unable to bathe" "wakes him up at night"	152	55 (59)
Diagnostic procedure	Procedure used in investigating the pain	"chest x-ray" "bloodwork"	268	70 (74)
Medication	Pharmacological pain treatment	"Oxycodone" "Tylenol"	593	77 (82)
Other treatment	Nonpharmacological methods used to alleviate the pain	"cortisol injection" "suggested CBT" (cognitive behavioral therapy)	789	83 (88)

Figure 5. The annotated chronic pain episodes, grouped by body location and with corresponding severity distributions. The sizes of the blue circles on the figure reflect the relative frequencies within the annotated cohort.

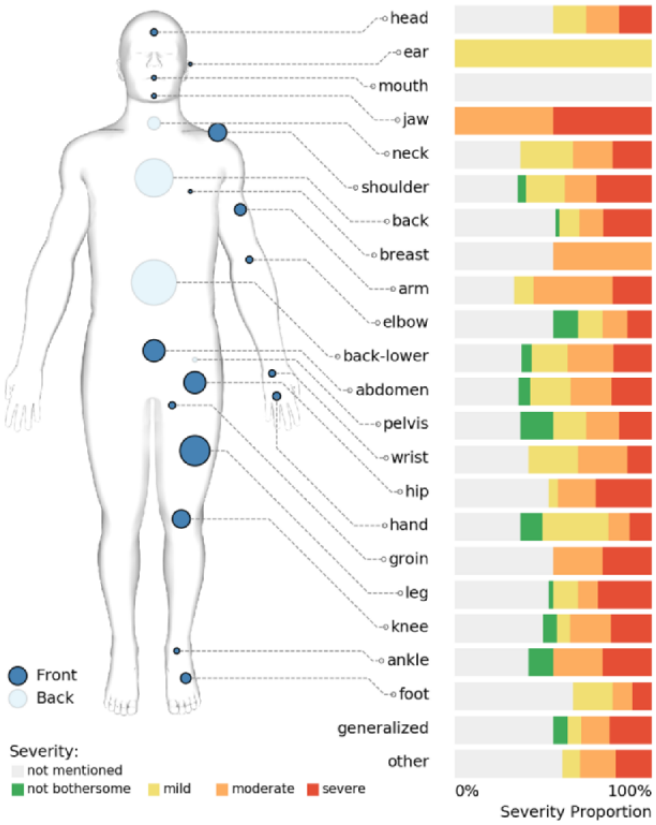


Figure 6. Illustration of an annotated episode exported into JavaScript Object Notation (JSON) format. Here, the mock-up patient had one episode of severe chronic pain in the back and the leg and received Percocet, which alleviated the pain. SCTID: SNOMED CT (Systematized Nomenclature Of Medicine–Clinical Terms) Identifier.


```
{
    "Patient_1":{
        "Patient_ID":123123,
        "Episode":{
            "Episode_1":{
                "Episode_ID":201409060,
                ”Episode_Identification_Date":20140901,
                 "Document_ID": 123456123,
                "End_Date":20150601,
                "Location":[
                    {
                        "back":{
                            "SCTID":77568009,
                            "spans":"23-27"
                        },
                        "leg":{
                            "SCTID":61685007,
                            "spans":"124-127"
                        }
                    }
                ],
                "Medication":[
                    {
                        "Percocet":{
                            "spans":"37-45",
                            "type":"prescription",
                            "order":"continuation_of_regimen",
                            "status":"present",
                            "effectiveness":"alleviates"
                        }
                    }
                ],
                "Severity":{
                    "spans":"72-80",
                    "severity":"severe_7_10"
                }
            }
        },
    }
}
```


The pain *location* and *severity* information are presented jointly in Figure 5. Back pain, including lower back pain, and leg pain were the most common types of chronic pain in this annotated cohort. Pain severity distributions varied by body location, but statistical testing was not performed due to the limited sample size for each body location.

Cause and *social and emotional effect* were two critical aspects of chronic pain that we sought to elicit from clinical text, as they are generally not available through structured data. Musculoskeletal issues were the leading cause of pain in the annotated episodes (61/94, 65%) (eg, degenerative arthritis or patellar tendinitis). Note that it was possible for a pain episode to have more than one cause (ie, mechanism). For example, an episode of musculoskeletal lower back pain could also be annotated as neuropathic if it included sciatica. The diverse social and emotional pain effects manifested both diminished efficacy and compromised quality of life.

...[the pain] wakes him up at night.

...affect his ability to study and perform lab work for schooling.

...has missed a couple of family functions due to pain.

We were also interested in the management aspects of chronic pain documented in the clinical notes. Opioids were found to be the most frequently used *medication* (49/94, 52%), followed by nonsteroidal anti-inflammatory drugs and acetaminophen. Physical and occupational therapies were the most frequently documented nonpharmacological *other treatment* (41/94, 44%), followed by analgesic injections and surgery. Another valuable aspect revealed in the narratives was sentiment around the treatments, as shown below. We did not attempt to assign numeric ratings to these sentiments, as sentiment analysis was not a major focus of this study.

...has been trying Tylenol without much relief.

...is wishing to discontinue her Cymbalta.

...consider gradually tapering down the Neurontin.

Interrater Disagreements

Most disagreements resulted from asymmetric annotation presence (ie, one annotator identified something the other annotator did not). In resolving these disagreements, we

determined that some represented underannotation and others overannotation.

Underspecified Concept Definition

Asymmetric presence of annotations usually emerged due to unclear or inadequate extensional definition of the attribute to be annotated. The typical scenario was that an annotator did not realize that an entity should be annotated. For example, "discomfort at rest" can be annotated as mild *severity*, but such qualification was not apparent unless specifically named in the guidelines. The inconsistency could be rooted in differential interpretation or domain literacy, which were compensated by the iteratively refined guidelines through patching of inclusion criteria as annotators gained more experience.

Overinference of Evidence

One annotator tended to extrapolate evidence beyond the text, which resulted in many disagreements on *location* and *social and emotional effect*. A typical example was that "carpal tunnel syndromes" got annotated as *wrist*. Although *wrist* could be inferred as the pain location, our discussion stipulated that the annotators should stick to the literal description (ie, syndrome instead of anatomic location) without overinference. This criterion was incorporated into the guidelines to discourage inferring evidence that is not explicitly mentioned.

Guideline Evolution

A summary of guideline evolution over the course of the study is available in Multimedia Appendix 1, section II, subsection 4. As the guideline was developed and used, some attributes were refined to make annotations more informative. For example, subcategories such as injury and trauma were added to the *cause* attribute to make the selections better fit the data. Other attributes were dropped due to scarce mentions in the corpus. Examples of dropped attributes include pain *trend*, which was intended to summarize whether pain was increasing, decreasing, or staying the same, and *referral*, which would identify a clinician referring a patient to another service.

Discussion

Principal Findings

Our episode-centered approach aligns closely with how clinicians identify and manage chronic pain: as an evolving and dynamic process. In some cases, including for aspects as central as the *cause* of chronic pain, clinicians exhibited changing attribution over time in their documentation, reflecting updated hypotheses as new information emerged (eg, from diagnostic interventions, progression of symptoms, or response to treatments). These time-varying considerations exemplify the complexity of accurately characterizing a chronic pain episode, which is more nuanced and multifaceted than simple concept extraction. Determining the episode boundaries can be nontrivial even for human annotators and relies on often imperfect or incomplete information in the text. It is also important to verify which pieces of evidence are credible and up to date whenever discrepancies in documentation are found. These challenges demand innovative solutions from the informatics community.

This study corroborated our hypothesis that rich information about chronic pain is available in clinical text and can be extracted with rigorous and standardized annotation. Some observations agreed with previous data, which have noted that back pain [15] was one of the most common causes of chronic pain, and opioids were a leading treatment choice in chronic pain [16]. Our annotation also produced novel data. For example, we found that chronic pain management is a multidisciplinary team effort that engages multiple medical and surgical departments, suggesting that optimal management will require active coordination that is attentive to specialty contexts.

Being the first clinical corpus dedicated to chronic pain, our experience could be considered prototypical with much room for improvement. Nonetheless, we believe the annotation guideline offers a solid starting point that can be referenced by other institutions with shared interest in abstracting chronic pain episodes from EHRs. We did not intend primarily to create or validate a corpus for machine learning, but we leave it to the potential users to determine how they want to leverage it. As a seed data set representing chronic pain annotations based on two health organizations, the corpus should benefit future work that has either a clinical or technical concentration (see Data Availability section for access information).

Limitations

To maximize the number of individual patients we could analyze, given limited annotator time, we prospectively decided to analyze notes from 6 months before to 2 years after the index diagnosis date. A distribution analysis in Table 5 finds that only 8 of the 94 (9%) episodes were censored at 24 months. As a result of this design decision and our stratified sampling method, our data should not be used to estimate average chronic pain episode length in a broader population.

Although our results summarize more than 3000 clinical notes, the sample size of 62 patients is not large. The studied cohort—patients from Midwestern United States with high access to health care—may not fully represent other populations within the United States. However, we believe many of our findings are representative of chronic pain and may serve as a useful comparison for populations with very different characteristics [17]. Moreover, the study population was derived from two health care systems, representing both academic and community practices. It is important to note that the ICD-9 codes used in this study may have excluded chronic pain of many other possible causes. Finally, as in any study using human annotators, the work reflected the knowledge and possibly subjective understanding of those who performed the task.

Future work can address these limitations by testing reproducibility in a different cohort and with an expanded cohort definition and time period. Another promising path we envision is to reconstruct a richly characterized trajectory for each chronic pain episode by interweaving pertinent evidence from multiple types of data.

Table 5. Distribution of episode-ending distance from the index diagnosis date.

Month after the index diagnosis date when episode ended	Episodes (N=94) , n (%)
0	21 (22)
2	3 (3)
4	5 (5)
6	3 (3)
8	5 (5)
10	3 (3)
12	5 (5)
14	1 (1)
16	9 (10)
18	5 (5)
20	14 (15)
22	12 (13)
24	8 (9)

Conclusions

To assess information about chronic pain in EHR notes, we performed systematic manual annotation on an age- and sex-stratified cohort of patients receiving health care at two health care systems between 2005 and 2015. An annotation guideline was iteratively refined to target key information about chronic pain episodes and associated attributes, such as *cause*, *location*, *severity*, and *medication*. A total of 3272 clinical notes were reviewed, and 94 episodes from 62 distinct patients were annotated. Summary statistics and qualitative analysis yielded insight into the characteristics of the cohort and their experiences. Clinical text was found to contain critical evidence for understanding the chronic pain trajectories of the patients. The episode-centered extraction captured a natural view of chronic pain while posing new challenges to potential automation.

Acknowledgments

The research was supported by the National Center for Advancing Translational Sciences (U01TR02062 and UL1TR002377 S2). We thank Dr Ahmad P Tafti for his valuable contributions during project discussions.

Authors' Contributions

JS and HL conceived the research idea. LC, MJ, WH, and JS developed the annotation guideline. LC and MJ performed the annotation in consultation with WH and JS for adjudication. LC, SF, and HH performed the data analysis and manuscript writing with supervision from JF. All authors contributed to discussion, interpretation of the data, and revision and approval of the final manuscript. JF and HL oversaw the execution of the study and serve as joint corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Guidelines on annotating chronic pain in clinical text.
[PDF File (Adobe PDF File), 369 KB - medinform_v8i11e18659_app1.pdf]

Multimedia Appendix 2
Subcategories of the annotated extrinsic chronic pain attributes.
[PDF File (Adobe PDF File), 92 KB - medinform_v8i11e18659_app2.pdf]

References

1. Dahlhamer J, Lucas J, Zelaya C, Nahin R, Mackey S, DeBar L, et al. Prevalence of chronic pain and high-impact chronic pain among adults - United States, 2016. MMWR Morb Mortal Wkly Rep 2018 Sep 14;67(36):1001-1006 [FREE Full text] [doi: 10.15585/mmwr.mm6736a2] [Medline: 30212442]

2. Duenas M, Ojeda B, Salazar A, Mico JA, Failde I. A review of chronic pain impact on patients, their social environment and the health care system. J Pain Res 2016 Jun;Volume 9:457-467. [doi: 10.2147/jpr.s105892]
3. Institute of Medicine. Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research. Washington, DC: The National Academies Press; 2011.
4. Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, et al. Chronic pain as a symptom or a disease: The IASP Classification of Chronic Pain for the International Classification of Diseases (ICD-11). Pain 2019 Jan;160(1):19-27. [doi: 10.1097/j.pain.0000000000001384] [Medline: 30586067]
5. Reuben DB, Alvanzo AA, Ashikaga T, Bogat GA, Callahan CM, Ruffing V, et al. National Institutes of Health Pathways to Prevention Workshop: The role of opioids in the treatment of chronic pain. Ann Intern Med 2015 Feb 17;162(4):295. [doi: 10.7326/m14-2775]
6. Saigh O, Triola MM, Link RN. Brief report: Failure of an electronic medical record tool to improve pain assessment documentation. J Gen Intern Med 2006 Feb;21(2):185-188. [doi: 10.1007/s11606-006-0256-z]
7. Tian TY, Zlateva I, Anderson DR. Using electronic health records data to identify patients with chronic pain in a primary care setting. J Am Med Inform Assoc 2013 Dec;20(e2):e275-e280 [FREE Full text] [doi: 10.1136/amiajnl-2013-001856] [Medline: 23904323]
8. Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, et al. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. J Am Med Inform Assoc 2013;20(5):898-905 [FREE Full text] [doi: 10.1136/amiajnl-2012-001076] [Medline: 23144336]
9. Dorflinger LM, Gilliam WP, Lee AW, Kerns RD. Development and application of an electronic health record information extraction tool to assess quality of pain management in primary care. Transl Behav Med 2014 Jun;4(2):184-189 [FREE Full text] [doi: 10.1007/s13142-014-0260-5] [Medline: 24904702]
10. Yim W, Tedesco D, Curtin C, Hernandez-Boussard T. Annotation of pain and anesthesia events for surgery-related processes and outcomes extraction. In: Proceedings of the Biomedical Natural Language Processing (BioNLP) Workshop. Stroudsburg, PA: Association for Computational Linguistics; 2017 Presented at: Biomedical Natural Language Processing (BioNLP) Workshop; August 4, 2017; Vancouver, Canada p. 200-205. [doi: 10.18653/v1/W17-2325]
11. Rocca WA, Yawn BP, St Sauver JL, Grossardt BR, Melton LJ. History of the Rochester Epidemiology Project: Half a century of medical records linkage in a US population. Mayo Clin Proc 2012 Dec;87(12):1202-1213. [doi: 10.1016/j.mayocp.2012.08.012] [Medline: 23199802]
12. St Sauver JL, Grossardt BR, Yawn BP, Melton LJ, Pankratz JJ, Brue SM, et al. Data resource profile: The Rochester Epidemiology Project (REP) medical records-linkage system. Int J Epidemiol 2012 Dec;41(6):1614-1624 [FREE Full text] [doi: 10.1093/ije/dys195] [Medline: 23159830]
13. Stubbs A. MAE and MAI: Lightweight annotation and adjudication tools. In: Proceedings of the 5th Linguistic Annotation Workshop. Stroudsburg, PA: Association for Computational Linguistics; 2011 Presented at: The 5th Linguistic Annotation Workshop; June 23-24, 2011; Portland, OR p. 129-133.
14. Hripcsak G. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005 Jan 31;12(3):296-298. [doi: 10.1197/jamia.m1733]
15. Gureje O, Von Korff M, Simon GE, Gater R. Persistent pain and well-being: A World Health Organization study in primary care. JAMA 1998 Jul 08;280(2):147-151. [doi: 10.1001/jama.280.2.147] [Medline: 9669787]
16. Martell BA, O'Connor PG, Kerns RD, Becker WC, Morales KH, Kosten TR, et al. Systematic review: Opioid treatment for chronic back pain: Prevalence, efficacy, and association with addiction. Ann Intern Med 2007 Jan 16;146(2):116-127. [doi: 10.7326/0003-4819-146-2-200701160-00006] [Medline: 17227935]
17. St Sauver JL, Grossardt BR, Leibson CL, Yawn BP, Melton LJ, Rocca WA. Generalizability of epidemiological findings and public health decisions: An illustration from the Rochester Epidemiology Project. Mayo Clin Proc 2012 Feb;87(2):151-160 [FREE Full text] [doi: 10.1016/j.mayocp.2011.11.009] [Medline: 22305027]

Abbreviations

EHR: electronic health record
IAA: interannotator agreement
ICD-9: International Classification of Diseases, Ninth Revision
IRB: Institutional Review Board
NLP: natural language processing
OMC: Olmsted Medical Center
REP: Rochester Epidemiology Project
SNOMED CT: Systematized Nomenclature Of Medicine–Clinical Terms

Edited by G Eysenbach; submitted 12.03.20; peer-reviewed by H Yu, A Louren, J Brenas; comments to author 12.06.20; revised version received 12.08.20; accepted 24.10.20; published 16.11.20.

Please cite as:
Carlson LA, Jeffery MM, Fu S, He H, McCoy RG, Wang Y, Hooten WM, St Sauver J, Liu H, Fan J
Characterizing Chronic Pain Episodes in Clinical Text at Two Health Care Systems: Comprehensive Annotation and Corpus Analysis
JMIR Med Inform 2020;8(11):e18659
URL: http://medinform.jmir.org/2020/11/e18659/
doi:10.2196/18659
PMID:33108311

Original Paper

Discovering the Context of People With Disabilities: Semantic Categorization Test and Environmental Factors Mapping of Word Embeddings from Reddit

Alejandro Garcia-Rudolph[1,2,3], PhD; Joan Saurí[1,2,3], PhD; Blanca Cegarra[1,2,3,4], MSc; Montserrat Bernabeu Guitart[1,2,3], MD

[1]Institut Guttmann Hospital de Neurorehabilitacio, Badalona, Spain

[2]Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain

[3]Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Badalona, Spain

[4]Universitat de Barcelona, Barcelona, Spain

Corresponding Author:
Alejandro Garcia-Rudolph, PhD
Institut Guttmann Hospital de Neurorehabilitacio
Camí de Can Ruti, s / n
Badalona,
Spain
Phone: 34 934977700
Email: alejandropablogarcia@gmail.com

Abstract

Background: The World Health Organization's International Classification of Functioning Disability and Health (ICF) conceptualizes disability not solely as a problem that resides in the individual, but as a health experience that occurs in a context. Word embeddings build on the idea that words that occur in similar contexts tend to have similar meanings. In spite of both sharing "context" as a key component, word embeddings have been scarcely applied in disability. In this work, we propose social media (particularly, Reddit) to link them.

Objective: The objective of our study is to train a model for generating word associations using a small dataset (a subreddit on disability) able to retrieve meaningful content. This content will be formally validated and applied to the discovery of related terms in the corpus of the disability subreddit that represent the physical, social, and attitudinal environment (as defined by a formal framework like the ICF) of people with disabilities.

Methods: Reddit data were collected from pushshift.io with the pushshiftr R package as a wrapper. A word2vec model was trained with the wordVectors R package using the disability subreddit comments, and a preliminary validation was performed using a subset of Mikolov analogies. We used Van Overschelde's updated and expanded version of the Battig and Montague norms to perform a semantic categories test. Silhouette coefficients were calculated using cosine distance from the wordVectors R package. For each of the 5 ICF environmental factors (EF), we selected representative subcategories addressing different aspects of daily living (ADLs); then, for each subcategory, we identified specific terms extracted from their formal ICF definition and ran the word2vec model to generate their nearest semantic terms, validating the obtained nearest semantic terms using public evidence. Finally, we applied the model to a specific subcategory of an EF involved in a relevant use case in the field of rehabilitation.

Results: We analyzed 96,314 comments posted between February 2009 and December 2019, by 10,411 Redditors. We trained word2vec and identified more than 30 analogies (eg, breakfast – 8 am + 8 pm = dinner). The semantic categorization test showed promising results over 60 categories; for example, s(A relative)=0.562, s(A sport)=0.475 provided remarkable explanations for low s values. We mapped the representative subcategories of all EF chapters and obtained the closest terms for each, which we confirmed with publications. This allowed immediate access (≤ 2 seconds) to the terms related to ADLs, ranging from apps "to know accessibility before you go" to adapted sports (boccia). For example, for the support and relationships EF subcategory, the closest term discovered by our model was "resilience," recently regarded as a key feature of rehabilitation, not yet having one unified definition. Our model discovered 10 closest terms, which we validated with publications, contributing to the "resilience" definition.

Conclusions: This study opens up interesting opportunities for the exploration and discovery of the use of a word2vec model that has been trained with a small disability dataset, leading to immediate, accurate, and often unknown (for authors, in many cases) terms related to ADLs within the ICF framework.

(JMIR Med Inform 2020;8(11):e17903) doi:10.2196/17903

KEYWORDS

disability; Reddit; social media; word2vec; semantic categorization; silhouette; activities of daily life; aspects of daily life; context; embeddings

Introduction

General Background

Natural Language Processing (NLP) is increasingly being integrated into several application domains. Google AI recently introduced BERT (Bidirectional Encoder Representations from Transformers) [1] to match search queries with more relevant results for optimizing Google searches. Facebook AI also achieved impressive breakthroughs, such as by tackling harmful or improper content by means of Whole Post Integrity Embeddings (WPIE) [2]. Other examples can be found in mobile apps, such as virtual assistants like Amazon's Alexa or Apple's Siri [3]. Application domains range from cultural heritage [4] to the identification of concepts and relationships in a body of research papers [5] or clinical decision support systems [6].

Words that occur in similar contexts tend to have similar meanings. This was likely first formulated in 1954 by Harris [7]. But the most famous statement of this principle came a few years later from linguist JR Firth: "You shall know a word by the company it keeps!" [8].

One of the strongest trends in NLP at the moment is the use of word embeddings, which are vectors whose relative similarities correlate with semantic similarity, building on the ideas of Harris and Firth.

The approval of the International Classification of Functioning, Disability, and Health (ICF) [9] by the World Health Assembly in May 2001 has marked a paradigm shift in the way health and disability are understood and measured [10]. The ICF conceptualizes disability not solely as a problem that resides in the individual but as a health experience that occurs in a *context* [11].

Disability and functioning are, according to the ICF model, outcomes of interactions between health conditions (diseases, disorders, and injuries) and contextual factors [9].

In spite of both sharing context as a key component, word embeddings have been scarcely applied in the field of disability, to the best of our knowledge.

In this paper, we hypothesize that social media can, indeed, link them. Word embeddings are usually learned from a general-purpose corpus; when it doesn't match the domain's vocabulary (including the same words or using words in the same senses), it is a problem that cannot simply be fixed with a lot of data. More data could just pull word contexts and representations towards generic, rather than domain-specific, values.

Our hypotheses in this paper are the following: (1) Such domain-specific values can be extracted from public domain-specific social media (2) in a sufficient number for the embedding to be relevant to the ICF model and (3) verifiable by sound theoretical semantic tests (4) consistent with state-of-the-art publications and (5) providing actionable knowledge to onfield specialists.

Social Media

Social media statistics from 2019 show that there are 3.2 billion social media users worldwide, and this number is growing [12].

Recent analyses remark that 42% of internet users take advantage of social media for health information, 32% of social media users in the United States share their health care experiences and family's struggle stories, and 29% search for health information via social media platforms to observe others' experiences with their diseases. Furthermore, 51% of those who live with a chronic disease have used the internet for information about health topics, such as details of a specific disease, medical procedures, drugs, medical devices, or health insurances [13].

Reddit

Social media users on platforms such as Reddit [14] tend to sharply contrast with similar groups that participate offline; for instance, people on Reddit are likely to discuss problems that they do not feel comfortable discussing face-to-face [15].

Another reason Reddit was chosen as a data source for this study is that the language of text posts is more structured than on other social media platforms such as Twitter [13].

As of 2019, Reddit's official statistics included 430 million monthly active users, 199 million posts, 1.7 billion comments, and around 14 billion views in a single month [16].

Reddit's core functionality is the sharing of text-based posts with others who may or may not be members of the site. The subforum function allows the creation of designated spaces for users to congregate and interact with each other over a shared interest. These subforums are called subreddits.

This Study

In the following subsections, we describe the specific characteristics and objectives of this study.

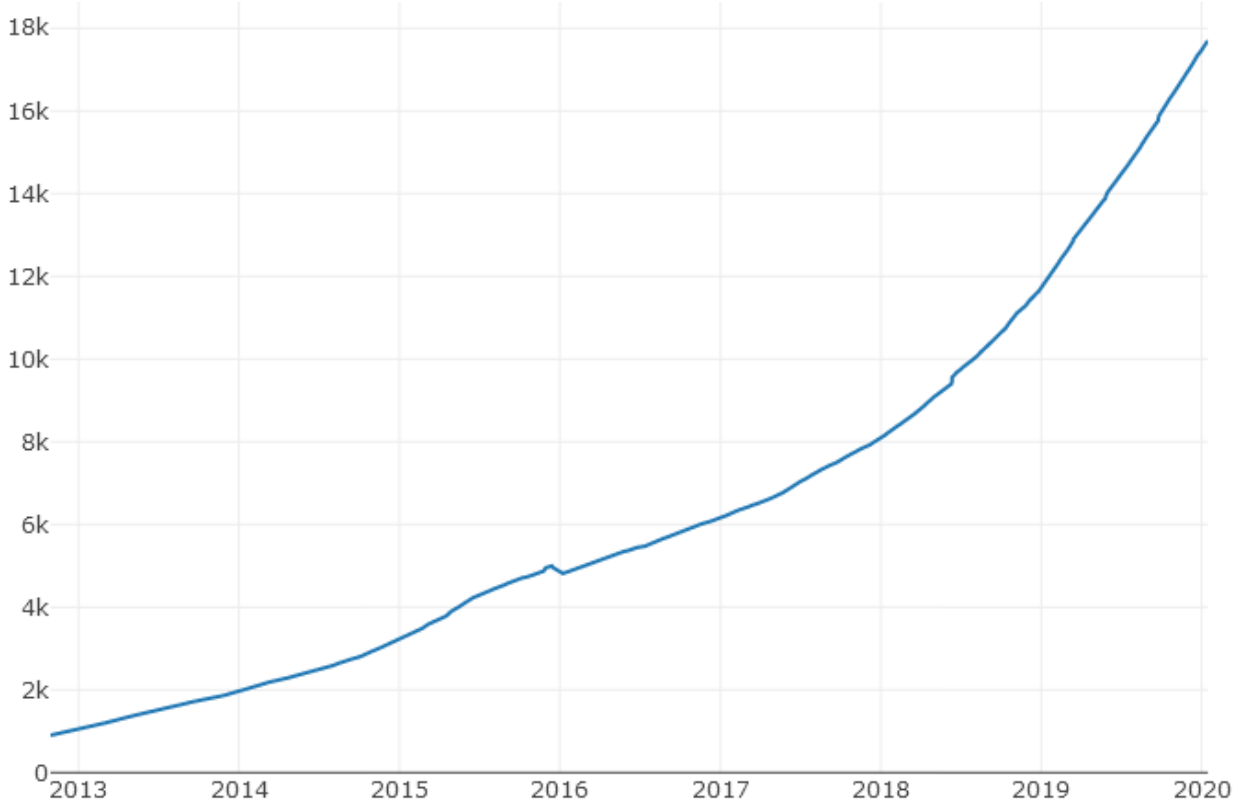
The Disability Subreddit

The data used for this study were extracted from the disability subreddit (containing news, resources, and perspectives pertaining to individuals with disabilities). It numbers 17,545 subscribers and 17 comments per day [17]. The evolution of

the number of subscribers since 2013 is shown in Figure 1. The disability subreddit was created on March 12, 2008.

The total number of posts and comments since 2008 are not shown in the Reddit official statistics; the 17 comments per day are, in fact, "the comments received on all its posts in a recent 24 hour measurement period. This number isn't averaged over time" [17]. A plot of the comments per day during the November 2018-December 2019 period can also be obtained from the official statistics [17] (Multimedia Appendix 1). Therefore, taking as a starting point the total number of subscribers (Figure 1) and the total number of comments during the last year, a rough estimation leads to about 100,000 comments during the 2009-2019 time period.

Figure 1. The number of subscribers in the disability subreddit by year.



Domain-Specific Values: Environmental Factors

According to the ICF, contextual factors represent the complete background of an individual's life and living [18]. They include 2 components: environmental factors (EF) and personal factors, which may have an impact on an individual with a health condition and that individual's health and health-related states.

In this study, we focus on the EF section of the ICF; by definition, EF make up the physical, social, and attitudinal environment in which people live and conduct their lives [9]. EF are organized into 5 chapters, each dealing with different and mutually exclusive aspects of the environment: (e1) products and technology, (e2) natural environment and human-made changes to the environment, (e3) support and relationships, (e4) attitudes, and (e5) services, systems, and policies.

Word2vec

Word2vec is the method used in this study for generating word embeddings. It creates an embedding (ie, numerical representations of words that help capture meaning, semantic relationships, and context) for text by using each word in a corpus to predict the words that usually surround it [19]. It consists of 2 neural network models: Continuous Bag of Words (CBOW) and Skip-gram. In both models, a window of predefined length is moved along the corpus, and in each step, the network is trained with the words inside the window to predict the word in the center of the window based on the surrounding words (CBOW) or to predict the contexts based on the central word (Skip-gram).

Therefore, word2vec creates word embeddings in which the semantic relationships between words are preserved. In this paper, we use the Skip-gram model, which shows better performance in semantic tasks [20].

Small Dataset: Twofold Validation of Word2vec Embedding

Word2vec methods have a distinct advantage in handling large datasets and have been trained with billions of tokens, as shown in the Google archive of the original Mikolov paper [21] (eg, the latest Wikipedia dumps 3 billion words, or the "one-billion-word language-modeling benchmark").

As it is a prediction-based model, it might be reasonable to expect that word2vec will produce very low-quality embeddings when trained with a small corpus.

Nevertheless, based on the high specificity of the disability subreddit, we hypothesize that the produced embeddings will

be of high semantic quality. In order to verify this, we will follow a twofold validation.

First, we will apply the semantic categorization test in order to measure the word2vec model's capabilities of representing semantic categories (such as vegetables, countries, fruits, and clothes). The original test (the Battig and Montague norm [22]) is composed of 53 categories with 10 words each. In order to measure how well the word *i* is grouped in relation to the other words in its semantic category, we will use silhouette coefficients [23].

Second, for each EF chapter (e1-e5), we will map meaningful word embeddings to representative categories of each chapter and refer to relevant publications to confirm their practical value.

Study Objectives

Specific aspects of disability (eg, depression) have been studied on social platforms such as Twitter [24] or support groups for Autism Spectrum Disorder on Facebook [25]. However, to our knowledge, no published study has examined social media content related to the environmental factors that make up the physical, social, and attitudinal environment in which people with disabilities live and conduct their lives.

Therefore, in this study, we aim to do the following: (1) extract all comments and submissions from the disability subreddit during the period under study (2009-2019); (2) train a word2vec model using disability subreddit comments as a training set, performing a preliminary validation using a subset of the original Mikolov paper analogies; (3) perform a semantic categorization test using an updated and expanded version of the Battig and Montague norm, with 65 categories; for each category, compute the silhouette coefficient of the model; (4) select representative subcategories addressing different aspects of daily life for each ICF chapter (e1-e5); for each subcategory, identify specific terms, t_i , t_j, extracted from their formal ICF definition, and run the word2vec model to generate the nearest semantic terms to t_i, and t_j; validate the obtained nearest semantic terms using relevant published literature; and (5) apply the results to a specific subcategory of a chapter involved in a relevant use case in the field of rehabilitation directly involved in daily living.

Methods

Data Collection

Reddit data were collected from pushshift.io [26] by the pushshift.io API (application programming interface). pushshift.io is a website that stores all publicly available Reddit submissions and comments, allowing researchers to collect and share Reddit datasets for research purposes, with extensive publications in related research (eg, Lama et al [27]). In this paper, we used the pushshiftr R package as a wrapper for the pushshift.io API [26].

Word2vec

For training the word2vec model, we used the wordVectors R package [28]. It implements the original C code for word2vec [20].

Semantic Distances

Given a vectorial representation of 2 words, their semantic similarity (S) was calculated using the cosine similarity measure between their respective vectorial representation, S(v1,v2). The semantic distances between 2 words, d(v1; v2), was calculated as 1 minus the semantic similarity, d(v1; v2)=1–S(v1; v2) [28].

Semantic Categorization Test

In this test, we measured the capabilities of the model to represent the semantic categories based on the Battig and Montague category norms, an invaluable tool for researchers in many fields, with a recent literature search revealing their use in over 1600 publications in more than 200 different journals [29].

In this study, we use Van Overschelde's [29] updated and expanded version of the Battig and Montague norms (expanded from 56 to 70 semantic categories).

In order to measure how well a word *i* is grouped in relation to the other words in its semantic category, we used the silhouette coefficients, *s(i)*.

a(i) is the mean distance of word *i* with all other words within the same category, and *b(i)* is the minimum mean distance of word *i* to any words within another category (ie, the mean distance to the neighboring category). In other words, silhouette coefficients measure how close a word is to other words within the same category compared to words of the closest category [23].

Results

Sample Description

Data were collected from 20,344 submissions and 96,314 comments from the disability subreddit.

Total comments were posted by 10,411 Redditors, and the total submissions were posted by 9658 Redditors.

The total number of different Redditors that have posted a submission or a comment is 15,072.

Considering that Reddit moderators remove a percentage of submissions (eg, for not following the Reddit posting guidelines), the number of different Redditors (15,072) is quite close to the 17,545 subscribers presented in Figure 1.

The first comment on the disability subreddit included in this analysis was published in February 2009, and the last was published in December 2019.

Note that these data were publicly accessible on Reddit and that no personally identifiable information is included in this study. The dataset is publicly available per request.

This paper focuses on the analysis of the 96,314 comments; Multimedia Appendix 1 presents further details on the retrieved data. For example, the most common 10 words are *people* (appearing 4231 times), *disability* (3619), *time* (3418), *disabled* (2479), *feel* (2457), *lot* (2433), *person* (2264), *life* (2162), *day*

(1841) and *job* (1811). Multimedia Appendix 1 also features a plot of the percentage of comments containing specific words (eg, *anger, hope, change, education*) by year, the fastest-growing words, and the words with the steepest increase in past years.

word2vec

We then ran the *train_word2vec* function of the wordVectors R package to train the model, with the following parameters: vectors=200, threads=4, window=12, iter=5, negative_samples=0.

We performed a preliminary validation using a reduced subset of the original Mikolov paper analogies [20]. We selected only some of those that might fit in our context; therefore, we did not include, for example, the well-known analogy

king–man+woman=queen

but we obtained promising results in a variety of analogies, such as

brother–sister+husband=wife(0.352) (cosine distance is shown in brackets)

This means that if, for our trained model, we execute the *nearest_to* function as follows:

nearest_to(model[["brother"]]-model[["sister"]]+model[["husband"]],5)

then we obtain "wife" as one of our top 5 nearest terms. We obtained several of them, with promising results. For example:

usa–ny+france=paris(0.666)

she–he+women=men(0.332)

doctor–hospital+teacher=school(0.599)

morning–woke+night=sleep(0.630)

girls–boys+women=men(0.474)

breakfast–8am+8pm=dinner(0.469)

cnn–news+netflix=streamer(0.515)

Semantic Categorization Test

For the first 5 words of each of the first 65 semantic categories of the updated version of the Battig and Montague norm [29], we calculated the silhouette coefficients. The complete list of the 325 words and silhouette calculation is presented in Multimedia Appendix 2.

Multimedia Appendix 3, which shows the silhouette coefficients for the first 60 semantic categories, displays promising results. The numbers identifying each category are those presented in the original norm; for example, the first row is "3. A relative" because this is the third semantic category presented in the norm.

When analyzing the lower silhouette scores, we identified remarkable reasons for the miscategorization of the terms. For example, regarding the first semantic category, "1. A precious stone," we selected *i*=diamond, and our model identified the category "43. A vegetable." To be closer to *i* than category 1, we tried to find out why broccoli is closer to diamond than, for example, a ruby, and we found that diamond is a well-known class of broccoli. Similar explanations can be found for the other miscategorizations; for example, for category "59. A liquid," we obtained *a(i)*=0.490 for *i*=water, but *b(i)* is lower [*b(i)*=0.381] because it is obtained for category "38. A non-alcoholic beverage." This is because water is the first term in category 38 and is included in the mean distance calculation for *b(i)*, (distance=0), but it is not included in the calculation of *a(i)* because the term *i* is not included in any calculation of *a(i)* (Multimedia Appendix 2). The same situation occurs with the category "20. An alcoholic beverage."

It is important to note the lowest silhouette we obtained, *s(i)*=-0.645, for *i*=tuna in the "52. A fish" category (Multimedia Appendix 3). In this case, *b(i)* is remarkably lower because the distances to category "43. A vegetable" are lower for all words in the category. This means that our model finds a closer relation between, for example, broccoli or lettuce with tuna (a healthy diet) than between tuna and shark, bass, beta, or cod, which are very difficult to relate only to fish names.

In spite of obtaining some low silhouette codes, the detailed analysis led to promising results. Table 1 shows the algebraic operations (eg, food+cholesterol) that we used to focus the closest terms onto the specific ICF subcategories, avoiding the miscategorization problems.

Table 1. Products and technology (e1).

2nd order code, 3rd order code, and algebraic operations	Closest terms (cosine distance)
e110 food and drugs	
e1100 food	
food+cholesterol	fruits (0.347), veggie (0.372), wheat (0.382), broccoli (0.413), oatmeal (0.413)
celiac+analysis	gluten (0.462), mitochondrial (0.510), coronary (0.513), psoriasis (0.519), genetic (0.526)
e1101 drugs	
drug+medicinal	marijuana(0.342), opioid(0.348), thc[a](0.433), benzos(0.498), cbd[b](0.451), well-butrin(0.504)
e115 assistive products	
e1151 specially designed	
voice+controlled	siri(0.400), dragon(0.406), alexa(0.437)
speech+app	voiceover(0.366), speechyfy (0.374), zoomtext(0.386)
e120 mobility transportation	
e1201 indoor outdoor	
motor+outdoor+power	powerchair (0.465), lightest (0.477), reclining(0.482), Invacare (0.502), wijits(0.506), pushrims(0.501)
transfer+indoor+device	hoist(0.434), stow(0.439), hoyer (0.449), ultralight(0.497)
taxi+wheelchair	uber (0.396), lyft (0.406), paratransit (0.420)
e140 culture recreation and sport	
e1401 adapted equipment	
adapted+sport	kayaking(0.389), archery (0.401), boccia (0.437),
e150 routing	
e1501 outdoor wayfinding	
wheelchair+apps	ableroad(0.541), wheelmap (0.551)

[a]thc: tetrahydrocannabinol.

[b]cbd: cannabidol.

Mapping to the ICF's Environmental Factors

Table 1 presents the closest terms to the representative subcategories of the products and technology chapter of the Environmental Factors. We present specific terms (eg, food and cholesterol, as illustrative examples) to show the potential of the model to discover relevant terms. In order to validate the results from Table 1, we went through the closest terms for each subcategory and identified recent publications and evidence.

The first terms for e1100, fruits and vegetables (veggie), have been extensively reported to be related to LDL (low-density lipoprotein) cholesterol [30], as well as wheat, oat [31], and broccoli [32].

We then considered other aspects related to food [eg, celiac disease (CD)]; we obtained *gluten* as the closest result, followed by *mitochondrial* (eg, reported by Picca et al [33]). The prevalence of the third term, *coronaryartery disease* (CAD), increases nearly twofold in patients with CD, as reported by Gajulapalli et al [34].

Ungprasert et al [35] demonstrated a significantly higher risk of CD among patients with psoriasis (fourth closest term) as well as genetic factors [36].

In relation to e1101 (drugs) cannabis, there is an increasing interest in the medical use of it (eg, in chronic pain, which is very common in disability) [37].

In relation to assistive products (eg, related to speech and voice), several state-of-the-art solutions were retrieved, such as Speechify, Zoomtext, Voiceover, Siri, Dragon, and Alexa.

In relation to e120, mobility transportation, desirable properties for outdoor transportation were retrieved, such as *lightest* or *reclining*, as well as top product providers (eg, Invacare and wijits). For indoor transportation, the closest terms were *hoist*, *stow*, and *hoyer*, as opposed to *powerchair*, which was retrieved for outdoors.

We then explored transportation services (taxi+wheelchair), and the closest terms were *uber, lyft*, and *paratransit*. It is important to remark that Uber [38] and Lyft provide specific disability policies [39].

In relation to e140, culture recreation and sport, several paralympic well-known ports emerged, such as *kayaking* or *archery*. Another close term was *boccia*, which is another (but perhaps less popular) Paralympic sport [40].

Finally, in relation to e1501, outdoor wayfinding, the wheelchair+app operation retrieved *ableroad* [41] and *wheelmap* [42] as closest terms.

Multimedia Appendix 2 includes a similar analysis for chapters e2-e5.

Rehabilitation Use Case: Resilience (ICF e398 Subcategory)

As shown in Multimedia Appendix 2, for the e398 support and relationships subcategory, the closest term identified by our model is *resilience*. Resilience has been recently regarded as a key feature of rehabilitation and living life well following a disability [43]. Evidence shows that resilience-based skills have multiple benefits once applied to people's lives (eg, a carry-over effect to other life domains). People do not have to be born resilient to become resilient; it can be improved with intentional practice [44].

It is unclear how well resilience or strategies to cultivate resilience are currently promoted as a component of rehabilitation programs. Furthermore, it does not yet have one unified definition, and research scholars have not decided on one specific understanding of what resilience means. Understanding resilience is an important component of building resilience [43].

Therefore, we used our model to identify its closest terms, and then we linked those terms to relevant publications. This is proposed as a straightforward example of how it can contribute to the understanding of a relevant term in rehabilitation. *Maturity*, *compassion*, *anger*, *resentment*, *grief*, *insecurity*, *contentment*, and *resenting* were identified by our model as the closest terms to *resilience*. Resilience and maturity have been extensively reported on (eg, by Davies et al [45]). The practice of compassion has been highlighted as an "essential component in nurturing resilience" [46]. As reported by Baldachino et al [47], the inability to cope effectively with anger may negatively impact a patient's physical and psychological well-being in the realm of resilience. Designing evidence-based interventions aimed at decreasing the negative impact of anger on resilience can be advanced by examining the potential mediation effect between anger and resilience [48].

As noted by Howard and Meichenbaum [49], resentment is a way of undermining resilience; resentment is a form of chronic, deep-seated anger. Holding onto resentment and not letting it go can have deleterious health effects and undermine the development of resilience. In relation to grief, it has also been related to resilience (eg, Bonano et al [50]).

As was presented in the EU Social Insecurities and Resilience Report [51], people who have low levels of social insecurities more often report high levels of resilience. In recent psychological research [52], contentment was also directly associated with both resilience and life satisfaction and mediated the relationship between these 2 aspects of well-being. Details and further analysis is presented in Multimedia Appendix 1.

Discussion

Principal Findings

In this paper, we proposed social media as the link between word embeddings and the ICF's environmental factors in a General Public License (GPL) framework (R-3.5.1). We applied a set of publicly available R libraries for collecting, model training, and analyzing Reddit public data. We trained a word2vec model using a small dataset and obtained encouraging results in king-queen–type analogies. Further, we obtained remarkable results in a standardized semantic categorization test. When mapping the discovered closest terms to representative subcategories of all ICF environmental factors, we verified them with scientific publications.

The obtained results open up interesting opportunities for exploration. Similarity isn't just a way of finding the nearest words; it is also a way of extracting items of a single class in every environmental factor that makes up the physical, social, and attitudinal environment in which people live and conduct their lives, ranging from apps "to know accessibility before you go" to adapted sports (boccia). Therefore, it can be thought of as a form of topic modeling; however, rather than letting the algorithm choose a fixed number of topics, it gives us the option of choosing the specific term (such as *resilience*) and how expansive we want the explored space to be.

Medical professionals are currently being encouraged to participate in social media, as remarked upon by Stukus [53]; even if a health care professional is not interested in engaging in social media, they must still be aware of the information people may be encountering online in order to provide anticipatory guidance in the clinical setting [47]. Therefore, this study also intends to be a step in that direction.

Limitations

The collected sample was not intended to be either representative or a comprehensive set of all comments posted by all persons with disabilities during the period under study. It includes all comments posted in Reddit's disability subreddit; we did not include comments from other subreddits addressing specific disability causes (eg, stroke) because, in this study, we follow the approach of studying disability in general (instead of any specific group or etiology), an approach also adopted in other recent literature reviews (eg, Hästbacka et al [54]). In addition, the ICF is grounded in the principle of universality, namely, that functioning and disability are applicable to all people, irrespective of health condition. The ICF is committed to the principle of parity, which states that functional status is not determined by background etiology or, in particular, by whether one has a physical rather than mental health condition [9].

Furthermore, including only 93,614 comments allowed us to verify our hypotheses.

We did not include submissions in our analysis; we collected them (20,344 in total), but we did not use them for model training because only 2745 of them were labeled as "questions"

while 1256 were labeled "articles and news," 438 of them were labeled "videos," and 13,570 were not labeled.

As presented in Multimedia Appendix 1, under the total number of comments by year, almost 80% of all comments took place during 2016-2019; therefore, the results of this study are weighted more strongly toward the recent years (2016-2019) rather than the early years (2009-2015) of the Disability subreddit.

Other relevant limitations to our study are related to the geographic location, spatial trajectory, or the time of day on which a comment was posted. As noted by Padilla et al [55] and Gore et al [56], such factors are relevant in social media. Spatiotemporal aspects are not controlled in our study; however, Reddit is most popular in the United States, with American users (representing 54% of Reddit's user base) far outnumbering those from other countries. Second to the United States, the UK represents the next highest share of Reddit traffic, at 8%, with Canada rounding out third, at 6.4%. Reddit is most popular amongst young adults between the ages of 25-34 years, which make up more than half of the site's users. Nevertheless, Reddit still draws in a large number of middle-aged users. A 2016 study found that people between the ages of 30-49 years accounted for 33% of the site's users, indicating that Reddit is a viable platform for reaching both young and middle-aged adults. Reddit is particularly popular among males, who make up more than two-thirds of the site's users [16].

Comparison with Prior Work

In medical applications, word2vec has recently been applied to larger datasets, such as 641,279 French health-related documents produced in a professional context (Dynomant et al [57]), 880,165 papers in a biomedical publication venue (Feng et al [58]), 1,451,413 abstracts for Adverse Drug Event Discovery Using Biomedical Literature (Tafti et al [59]), and 1,749,870 reviews in Online Doctor Reviews [60].

Reddit has recently been used as a data source for studies in chronic diseases, such as Foufi et al's [13] analysis of 17,624 text posts for entity and relation extraction, employing the PKDE4J tool. Sharma et al [61] performed a qualitative analysis of Reddit discussions regarding motivations and limitations associated with vaping among people with mental illness, a thematic analysis that included 3263 comments from 133 discussion threads.

In a small corpus (37,000 and 140,000 documents) using a semantic categorization test, word2vec was applied in analyzing and disambiguating the content of dreams [62]; this research field addresses questions such as "how do gender, cultural background, and waking-life experiences shape dream content?".

Facebook groups, discussion forums, and chat rooms were recently analyzed [63] to explore and compare the interactions and connections among online support groups to better understand how people with disabilities were utilizing different social networks to facilitate communication interchange. They concluded that through participation on different platforms, persons with disabilities are able to provide and receive social support in various ways, without the barriers and constraints often experienced by this population.

Our approach is completely different in that we provide a tool for discovering terms of interest (in this paper, we applied it to the ICF environmental factors, but it could also be applied to, for example, the ICF's body functions and structures). For example, regarding e5 services and associations, in using the 3 terms *hear+association+kid*, we obtained *depaul* as one of the closest results. The DePaul School for Hearing and Speech teaches children who are deaf or hard of hearing to listen and speak without sign language [64]. Another obtained close term is *saskatoon*, a new early learning pilot program in Saskatoon that impacts preschoolers who are deaf or hard of hearing by breaking down communication barriers [65].

Therefore, it can be addressed to a wide variety of involved stakeholders besides people with disabilities themselves, such as informal caregivers, health care professionals, and private or public associations.

Conclusions

This study explored the ability of word2vec to extract the main factors affecting the lives of people with disabilities within the ICF framework from a small dataset, showing promising results.

Our results open up interesting opportunities for exploration and discovery. Similarity is revealed as not only a way of finding the nearest words but also a way of extracting out items related to specific elements. Therefore, it can be thought of as a form of topic modeling, where users can focus on a particular term in-breath or in-depth.

Acknowledgments

This research was partially funded by PARTICIPA Barriers and Facilitators to Social Participation and PRECISE4Q Personalised Medicine by Predictive Modeling in Stroke for Better Quality of Life H2020, grant number 777107.

Authors' Contributions

AGR and JSR conceived the study. AGR, JSR, and BCR collected, selected, and cleaned the data. AGR and BCD analyzed the data. AGR and BCD drafted the initial manuscript. JSR and MBG revised the manuscript critically for important intellectual content and approved the final manuscript. AGR, JSR, and MBG received funding for the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Word trends, closest terms, environmental factors, and resilience.
[DOCX File , 212 KB - medinform_v8i11e17903_app1.docx]

Multimedia Appendix 2
Semantic Categorization Test.
[XLSX File (Microsoft Excel File), 427 KB - medinform_v8i11e17903_app2.xlsx]

Multimedia Appendix 3
The 60 semantic categories ordered by silhouette value.
[DOCX File , 13 KB - medinform_v8i11e17903_app3.docx]

References

1. Clark K, Khandelwal U, Levy O, Manning CD. What Does BERT Look at? An Analysis of BERT's Attention. : The Association for Computational Linguistics; 2019 Presented at: The BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP at ACL 2019; August 2019; Florence, Italy. [doi: 10.18653/v1/w19-4828]
2. MacAvaney S, Yao HR, Yang E, Russell K, Goharian N. Hate speech detection: Challenges and solutions. PLOS ONE 2019;14 (8) [FREE Full text] [doi: 10.1371/journal.pone.0221152]
3. Ni Loideain N, Adams R. From Alexa to Siri and the GDPR: The Gendering of Virtual Personal Assistants and the Role of EU Data Protection Law. SSRN Journal 2018. [doi: 10.2139/ssrn.3281807]
4. Sporleder C. Natural Language Processing for Cultural Heritage Domains. Language and Linguistics Compass 2010;4(10). [doi: 10.1111/j.1749-818x.2010.00230.x]
5. Diallo S, Gore R, Padilla J, Lynch C. An Overview of Modeling and Simulation using Content Analysis. Scientometrics 2015. [doi: 10.1007/s11192-015-1578-6]
6. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct 26;42(5):760-772 [FREE Full text] [doi: 10.1016/j.jbi.2009.08.007] [Medline: 19683066]
7. Harris ZS. Distributional Structure. New York: Verlag; 1954.
8. Firth JR. A synopsis of linguistic theory 1930–1955. In Studies in Linguistic Analysis. Oxford: Palmer; 1957.
9. World Health Organization Geneva. Towards a Common Language for Functioning, Disability and Health. The International Classification of Functioning, Disability and Health. 2002. URL: https://www.who.int/classifications/icf/icfbeginnersguide.pdf [accessed 2020-01-10] [WebCite Cache ID www.who.int/classifications/icf/icfbeginnersguide.pdf]
10. Kostanjsek N. Use of The International Classification of Functioning, Disability and Health (ICF) as a conceptual framework and common language for disability statistics and health information systems. BMC Public Health 2011 May 31;11 Suppl 4(2-3):S3-162 [FREE Full text] [doi: 10.1186/1471-2458-11-S4-S3] [Medline: 21624189]
11. Quah SR, editor. International Encyclopedia of Public Health, 2nd Ed. Cambridge: Academic Press; Nov 03, 2016.
12. Oberlo. 10 Social Media Statistics You Need to Know. URL: https://www.oberlo.com/blog/social-media-marketing-statistics [accessed 2020-01-10] [WebCite Cache ID www.oberlo.com/blog/social-media-marketing-statistics]
13. Foufi V, Timakum T, Gaudet-Blavignac C, Lovis C, Song M. Mining of Textual Health Information from Reddit: Analysis of Chronic Diseases With Extracted Entities and Their Relations. J Med Internet Res 2019 Jun 13;21(6) [FREE Full text] [doi: 10.2196/12876] [Medline: 31199327]
14. reddit. 2002. URL: www.reddit.com [accessed 2020-01-10] [WebCite Cache ID www.reddit.com]
15. Johnson GJ, Ambrose PJ. Neo-tribes: the power and potential of online communities in health care. Commun. ACM 2006 Jan;49(1):107-113 [FREE Full text] [doi: 10.1145/1107458.1107463]
16. Reddit stats. URL: https://redditblog.com/2019/12/04/reddits-2019-year-in-review/ [accessed 2020-01-10] [WebCite Cache ID https://redditblog.com/2019/12/04/reddits-2019-year-in-review/]
17. Disability subreddit. URL: https://subredditstats.com/r/disability [accessed 2020-01-10] [WebCite Cache ID https://subredditstats.com/r/disability]
18. Grotkamp S, Cibis W, Nüchtern E, von Mittelstaedt G, Seger W. Personal Factors in the International Classification of Functioning, Disability and Health: Prospective Evidence. The Australian Journal of Rehabilitation Counselling 2012 Jul 04;18(1):1-24. [doi: 10.1017/jrc.2012.4]
19. Allem JP, Dharmapuri L, Unger JB, Cruz TB. Characterizing JUUL-related posts on Twitter. Drug Alcohol Depend 2018 Sep 01;190:1-5 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.05.018] [Medline: 29958115]
20. Mikolov T, Corrado G, Chen K, Dean J. Efficient Estimation of Word Representations in Vector Space. In: Proceedings of the International Conference on Learning Representations. 2013 Presented at: International Conference on Learning Representations; May 2nd to May 4th 2013; Scottsdale, Arizona p. 1-12.
21. Google word2vec. URL: https://code.google.com/archive/p/word2vec/ [accessed 2020-01-10] [WebCite Cache ID https://code.google.com/archive/p/word2vec/]

22. Battig WF, Montague WE. Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. Journal of Experimental Psychology 1969 Jun 13;80(3, Pt.2):1-46. [doi: 10.1037/h0027577]
23. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987 Nov 13;20(6):53-65. [doi: 10.1016/0377-0427(87)90125-7]
24. Leis A, Ronzano F, Mayer MA, Furlong LI, Sanz F. Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. J Med Internet Res 2019 Jun 27;21(6):e14199-e14101 [FREE Full text] [doi: 10.2196/14199] [Medline: 31250832]
25. Mustafa HR, Short M, Fan S. Social Support Exchanges in Facebook Social Support Group. Procedia - Social and Behavioral Sciences 2015 May;185:346-351. [doi: 10.1016/j.sbspro.2015.03.449]
26. pushshift.io: API Documentation. URL: https://pushshift.io/api-parameters/ [accessed 2020-01-10] [WebCite Cache ID https://pushshift.io/api-parameters/]
27. Lama Y, Hu D, Jamison A, Quinn SC, Broniatowski DA. Characterizing Trends in Human Papillomavirus Vaccine Discourse on Reddit (2007-2015): An Observational Study. JMIR Public Health Surveill 2019 Mar 27;5(1):e12480-e12113 [FREE Full text] [doi: 10.2196/12480] [Medline: 30916662]
28. wordVectors. URL: https://github.com/bmschmidt/wordVectors [accessed 2020-01-10] [WebCite Cache ID https://github.com/bmschmidt/wordVectors]
29. Van Overschelde J, Rawson K, Dunlosky J. Category norms: An updated and expanded version of the Battig and Montague (1969) norms. Journal of Memory and Language 2004 Apr;50(3):289-335. [doi: 10.1016/j.jml.2003.10.003]
30. Djoussé L, Arnett DK, Coon H, Province MA, Moore LI, Ellison RC. Fruit and vegetable consumption and LDL cholesterol: the National Heart, Lung, and Blood Institute Family Heart Study. Am J Clin Nutr 2004 Feb;79(2):213-217. [doi: 10.1093/ajcn/79.2.213] [Medline: 14749225]
31. Momenizadeh A, Heidari R, Sadeghi M, Tabesh F, Ekramzadeh M, Haghighatian Z, et al. Effects of oat and wheat bread consumption on lipid profile, blood sugar, and endothelial function in hypercholesterolemic patients: A randomized controlled clinical trial. ARYA Atheroscler 2014 Sep;10(5):259-265 [FREE Full text] [Medline: 25477983]
32. Armah CN, Derdemezis C, Traka MH, Dainty JR, Doleman JF, Saha S, et al. Diet rich in high glucoraphanin broccoli reduces plasma LDL cholesterol: Evidence from randomised controlled trials. Mol Nutr Food Res 2015 May 04;59(5):918-926 [FREE Full text] [doi: 10.1002/mnfr.201400863] [Medline: 25851421]
33. Picca A, Riezzo G, Lezza AMS, Clemente C, Pesce V, Orlando A, et al. Mitochondria and redox balance in coeliac disease: A case-control study. Eur J Clin Invest 2018 Feb 04;48(2):1-24. [doi: 10.1111/eci.12877] [Medline: 29243228]
34. Gajulapalli RD, Pattanshetty DJ. Risk of coronary artery disease in celiac disease population. Saudi J Gastroenterol 2017 Sep;23(4):253-258 [FREE Full text] [doi: 10.4103/sjg.SJG_616_16] [Medline: 28721980]
35. Ungprasert P, Wijarnpreecha K, Kittanamongkolchai W. Psoriasis and Risk of Celiac Disease: A Systematic Review and Meta-analysis. Indian J Dermatol 2017;62(1):41-46. [doi: 10.4103/0019-5154.198031] [Medline: 28216724]
36. Farina F, Picascia S, Pisapia L, Barba P, Vitale S, Franzese A, et al. HLA-DQA1 and HLA-DQB1 Alleles, Conferring Susceptibility to Celiac Disease and Type 1 Diabetes, are More Expressed Than Non-Predisposing Alleles and are Coordinately Regulated. Cells 2019 Jul 19;8(7):2020-2001 [FREE Full text] [doi: 10.3390/cells8070751] [Medline: 31331105]
37. Poli P, Crestani F, Salvadori C, Valenti I, Sannino C. Medical Cannabis in Patients with Chronic Pain: Effect on Pain Relief, Pain Disability, and Psychological aspects. A Prospective Non randomized Single Arm Clinical Trial. Clin Ter 2018;169(3):e102-e107 [FREE Full text] [doi: 10.7417/T.2018.2062] [Medline: 29938740]
38. UBER Accessibility. URL: https://www.uber.com/us/en/about/accessibility/ [accessed 2020-01-10] [WebCite Cache ID https://www.uber.com/us/en/about/accessibility/]
39. Lyft Wheelchair policy. URL: https://help.lyft.com/hc/en-us/articles/115012926827-Wheelchair-Policy [accessed 2020-01-10]
40. Sports: Boccia. URL: https://www.disabledsportsusa.org/sport/boccia/ [accessed 2020-01-10]
41. iaccessiblity.com. AbleRoad. URL: https://www.iaccessibility.com/apps/general/index.cgi/product?ID=3 [accessed 2020-01-10]
42. Wheelmap. URL: https://news.wheelmap.org/en/wheelmap-org-750000-places-in-seven-years/ [accessed 2020-01-10]
43. Stuntzner S, Dalton J, Umeasiegbu V, MacDonald A, Mercado F. Resilience and Disability: Consideration and Integration of Resilience Training in Undergraduate Rehabilitation Service Programs. Journal of Applied Rehabilitation Counseling 2018 Dec 01;49(4):5-13. [doi: 10.1891/0047-2220.49.4.5]
44. Newman R. APA's resilience initiative. Professional Psychology: Research and Practice 2005 Jun;36(3):227-229. [doi: 10.1037/0735-7028.36.3.227]
45. Davies M. Disability and Impairment: Working with Children and Families. British Journal of Social Work 2007 Feb 01;38(8):1656-1657. [doi: 10.1093/bjsw/bcn155]
46. Marini I, Milington M, Graf NM. Psychosocial Aspects of Disability, 2nd Ed. New York: Springer Publishing Company; 2017.
47. Baldacchino DR. Student nurses' personality traits and the nursing profession: part 1. Br J Nurs 2012;21(7):419-425. [doi: 10.12968/bjon.2012.21.7.419] [Medline: 22585020]

48. Wu W, Chang J, Tsai S, Liang S. Assessing Self-concept as a Mediator Between Anger and Resilience in Adolescents With Cancer in Taiwan. Cancer Nursing 2018 Mar 27;41(3):210-217. [doi: 10.1097/NCC.0000000000000512]
49. Meichenbaum D. Roadmap to Resilience: A Guide for Military, Trauma Victims and Their Families. United Kingdom: CROWN HOUSE PUB LTD; Sep 15, 2012.
50. Bonanno G, Wortman C, Lehman D, Tweed R, Haring M, Sonnega J, et al. Resilience to loss and chronic grief: A prospective study from preloss to 18-months postloss. Journal of Personality and Social Psychology 2002;83(5):1150-1164. [doi: 10.1037/0022-3514.83.5.1150]
51. , editor. Social insecurities and resilience. Luxembourg: Publications Office of the European Union; 2018.
52. Gerson MW. Spirituality, Social Support, Pride, and Contentment as Differential Predictors of Resilience and Life Satisfaction in Emerging Adulthood. Psychology 2018 Apr;09(03):485-517. [doi: 10.4236/psych.2018.93030]
53. Stukus DR, Patrick MD, Nuss K. Social Media for Medical Professionals: Strategies for Successfully Engaging in an Online World. New York: Springer; May 13, 2013.
54. Hästbacka E, Nygård M, Nyqvist F. Barriers and facilitators to societal participation of people with disabilities: A scoping review of studies concerning European countries. European Journal of Disability Research 2016 Jul;10(3):201-220. [doi: 10.1016/j.alter.2016.02.002]
55. Padilla J, Kavak H, Lynch C, Gore R, Diallo S. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. PLoS One 2018;13(6):e0198857. [doi: 10.1371/journal.pone.0198857] [Medline: 29902270]
56. Gore R, Diallo S, Padilla J. You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. PLoS One 2015 Sep;10(9):e0133505-e0133565. [doi: 10.1371/journal.pone.0133505] [Medline: 26332588]
57. Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, et al. Word Embedding for the French Natural Language in Health Care: Comparative Study. JMIR Med Inform 2019 Jul 29;7(3):e12310-e12926 [FREE Full text] [doi: 10.2196/12310] [Medline: 31359873]
58. Feng X, Zhang H, Ren Y, Shang P, Zhu Y, Liang Y, et al. The Deep Learning-Based Recommender System "Pubmender" for Choosing a Biomedical Publication Venue: Development and Validation Study. J Med Internet Res 2019 May 24;21(5):e12957 [FREE Full text] [doi: 10.2196/12957] [Medline: 31127715]
59. Tafti AP, Badger J, LaRose E, Shirzadi E, Mahnke A, Mayer J, et al. Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure. JMIR Med Inform 2017 Dec 08;5(4):e51 [FREE Full text] [doi: 10.2196/medinform.9170] [Medline: 29222076]
60. Rivas R, Montazeri N, Le N, Hristidis V. Automatic Classification of Online Doctor Reviews: Evaluation of Text Classifier Algorithms. J Med Internet Res 2018 Nov 12;20(11):e11141 [FREE Full text] [doi: 10.2196/11141] [Medline: 30425030]
61. Sharma R, Wigginton B, Meurk C, Ford P, Gartner CE. Motivations and Limitations Associated with Vaping among People with Mental Illness: A Qualitative Analysis of Reddit Discussions. Int J Environ Res Public Health 2016 Dec 22;14(1):751 [FREE Full text] [doi: 10.3390/ijerph14010007] [Medline: 28025516]
62. Altszyler E, Ribeiro S, Sigman M, Fernández Slezak D. Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. Conscious Cogn 2017 Nov;56(7):178-187.
63. Stetten NE, LeBeau K, Aguirre MA, Vogt AB, Quintana JR, Jennings AR, et al. Analyzing the Communication Interchange of Individuals With Disabilities Utilizing Facebook, Discussion Forums, and Chat Rooms: Qualitative Content Analysis of Online Disabilities Support Groups. JMIR Rehabil Assist Technol 2019 Sep 30;6(2):e12667 [FREE Full text] [doi: 10.2196/12667] [Medline: 31573937]
64. DePaul School for Hearing & Speech. URL: https://www.depaulhearingandspeech.org/ [accessed 2020-01-10]
65. Stillger N. New program in Saskatoon having positive impact on deaf and hard of hearing preschoolers. Global News. 2018 Dec 11. URL: https://globalnews.ca/news/4752539/preschoolers-deaf-hard-of-hearing-saskatoon-regina/ [accessed 2020-01-10] [WebCite Cache ID https://globalnews.ca/news/4752539/preschoolers-deaf-hard-of-hearing-saskatoon-regina/]

Abbreviations

ADLs: aspects of daily living
API: application program interface
BERT: bidirectional encoder representations from transformers
CBD: cannabidiol
CBOW: continuous bag of words
EF: environmental factors
ICF: International Classification of Functioning Disability and Health
LDL: low-density lipoprotein
NLP: natural language processing
THC: tetrahydrocannabinol
WPIE: whole post integrity embeddings

Edited by G Eysenbach; submitted 20.01.20; peer-reviewed by R Gore, M Bjelogrlic; comments to author 10.02.20; revised version received 17.04.20; accepted 19.04.20; published 20.11.20.

Please cite as:
Garcia-Rudolph A, Saurí J, Cegarra B, Bernabeu Guitart M
Discovering the Context of People With Disabilities: Semantic Categorization Test and Environmental Factors Mapping of Word Embeddings from Reddit
JMIR Med Inform 2020;8(11):e17903
URL: http://medinform.jmir.org/2020/11/e17903/
doi:10.2196/17903
PMID:33216006

Original Paper

Identification of Adverse Drug Event–Related Japanese Articles: Natural Language Processing Analysis

Shogo Ujiie[1], BA; Shuntaro Yada[1], PhD; Shoko Wakamiya[1], PhD; Eiji Aramaki[1], PhD

Nara Institute of Science and Technology, Nara, Japan

Corresponding Author:
Eiji Aramaki, PhD
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma
Nara, 630-0192
Japan
Phone: 81 743 72 5250
Email: aramaki@is.naist.jp

Abstract

Background: Medical articles covering adverse drug events (ADEs) are systematically reported by pharmaceutical companies for drug safety information purposes. Although policies governing reporting to regulatory bodies vary among countries and regions, all medical article reporting may be categorized as precision or recall based. Recall-based reporting, which is implemented in Japan, requires the reporting of any possible ADE. Therefore, recall-based reporting can introduce numerous false negatives or substantial amounts of noise, a problem that is difficult to address using limited manual labor.

Objective: Our aim was to develop an automated system that could identify ADE-related medical articles, support recall-based reporting, and alleviate manual labor in Japanese pharmaceutical companies.

Methods: Using medical articles as input, our system based on natural language processing applies document-level classification to extract articles containing ADEs (replacing manual labor in the first screening) and sentence-level classification to extract sentences within those articles that imply ADEs (thus supporting experts in the second screening). We used 509 Japanese medical articles annotated by a medical engineer to evaluate the performance of the proposed system.

Results: Document-level classification yielded an F1 of 0.903. Sentence-level classification yielded an F1 of 0.413. These were averages of fivefold cross-validations.

Conclusions: A simple automated system may alleviate the manual labor involved in screening drug safety–related medical articles in pharmaceutical companies. After improving the accuracy of the sentence-level classification by considering a wider context, we intend to apply this system toward real-world postmarketing surveillance.

(JMIR Med Inform 2020;8(11):e22661) doi:10.2196/22661

KEYWORDS

adverse drug events; medical informatics; natural language processing; pharmacovigilance

Introduction

Background

According to the World Health Organization, an adverse drug event (ADE) is any untoward occurrence that may present during treatment with a pharmaceutical product but is not necessarily causally related to the treatment [1]. According to a survey conducted by Howard et al [2], ADEs are responsible for approximately 3.7% of all hospital admissions worldwide. This issue has been addressed by institutional premarketing and postmarketing drug safety surveillance. However, postmarketing measures play a more important role than premarketing clinical trials, as postmarketing measures can also detect infrequent reactions, long-term effects, and drug–drug/drug–food interactions [3]. A major source of postmarketing surveillance is spontaneous or voluntary reporting of suspected adverse reactions by clinicians, pharmacists, and the pharmaceutical industry. However, owing to the high volume of incoming "signals," the identification of even a few credible reports is labor-intensive [4]. Hence, the development of an automated system that determines and classifies the relative importance of clinical ADE-related reports would be considerably beneficial.

Existing automation research targets different source materials, reflecting the wide range of signals processed by real-world postmarketing surveillance. These inputs include electronic health records [5,6], patient reports [7], medical articles [8,9], and social media posts [10,11]. This study focuses on medical articles as they comprise a substantial portion of postmarketing surveillance in many countries. Pharmaceutical companies voluntarily send reports based on medical articles to regulatory bodies [4].

Policies governing the reporting of ADEs to regulatory bodies vary among countries and regions [12]. In general, reporting may be precision or recall based. In the former approach, implemented in the United States [13] and the European Union [14], suspected serious adverse drug reactions (ADRs) are rapidly reported. Serious ADRs correspond to certain ADEs for which reasonable causal relationships between events and drugs are suspected or confirmed. In recall-based reporting, any possible ADE must be reported immediately. ADEs include cases where a causal relationship between drugs and harmful events cannot be ruled out [15]. Nevertheless, this strategy can introduce numerous false negatives or substantial amounts of noise. Processing a large volume of reports on all possible ADEs greatly increases the manual classification burden. Overall, recall-based reporting is very difficult to accomplish using limited human labor.

Japan has adopted and implemented recall-based pharmacovigilance [15]. Its main information source is spontaneous reporting from pharmaceutical companies, and the basis of these reports is medical articles. This process usually consists of a first (initial) screening followed by a second screening. In the first screening, medical articles are manually classified and prioritized. For example, if an article mentions fatal or lethal ADEs, it receives top priority. The second in-depth screening is performed by medical experts and assesses the merit of the reported ADEs. In the Japanese pharmaceutical company involved in this survey, thousands of articles must be monitored annually, and each report requires at least 10 minutes to evaluate. This process incurs a significant labor cost. Moreover, the criteria used in the first screening may be subjective (and thus vary considerably according to the person conducting it). Consequently, the surveillance process may be unnecessarily delayed.

Objectives

To address Japanese pharmacovigilance, we have developed an automated system that replaced the first screening by extracting ADE-containing articles. For the second screening, we also enlisted the services of medical experts to identify ADE-suggesting sentences in the articles. Our system combines both document- and sentence-level classification models. It classifies Japanese medical articles to extract those that contain references to ADEs and then uses them as candidates for the second screening (*ADE-containing article extraction*). It also classifies the ADE-suggesting sentences that must be scrutinized by medical experts (*ADE-suggesting sentence extraction*).

To this end, we implemented natural language processing (NLP) techniques. Our system consists of simple machine learning methods that are easily applied and managed in-house by pharmaceutical companies. Targeting Japanese medical articles also offers insights into an effective management approach for papers written in non-English languages with few linguistic resources within the medical domain.

- To support postmarketing surveillance in Japan where recall-based reporting is adopted for drug safety, we built an automated system identifying ADE-containing medical articles and the ADE-suggesting sentences therein to improve interpretability.
- Our proposed models classified the ADE-containing articles at a 0.903 F1 score and ADE-suggesting sentences at a 0.413 F1 score based on a manually annotated test set of Japanese medical articles.
- We developed an effective automated system based on relatively simple models. It can be easily implemented and managed in-house by pharmaceutical companies. In addition, our system may be readily expanded to classify papers written in non-English languages.

Methods

Materials

Japanese medical articles used for postmarketing surveillance duty in a Japanese pharmaceutical company were provided for subsequent analysis. The majority of the articles were related to a select range of drugs surveyed by the company, but were not limited to specific clinical areas or diseases. The frequency with which each drug appears in the data is reported in Multimedia Appendix 1. The articles were randomly sampled, and text data were extracted from PDF documents by optical character recognition (OCR) using WinReader PRO version 15.0 (NTTDATA NJK Corporation) [16]. Articles written in 2 or more columns were excluded as the OCR software had difficulty recognizing text in this format. Subsequent filtering generated 509 medical articles. Certain symbols such as “$”and “^” that do not normally appear in Japanese medical articles were removed.

After preprocessing, all sentences were filtered on the basis of the appearance of ADEs. These were judged according to the following criteria:

1. An adverse event was mentioned after a drug prescription; or
2. The author explicitly mentioned the occurrence of a suspicious ADE.

Matched sentences were labeled as *ADE suggesting*. This determination was made by considering multiple sentences. Hence, the ADE-suggesting designation may have spanned several sentences.

Medical articles containing any ADE-suggesting sentences were designated *ADE containing*. Here, 300/509 articles (58.9%) were labeled ADE containing. Table 1 shows the average number of sentences and characters per medical article in the respective ADE-containing articles.

Table 1. Average number of sentences and characters per medical article.

Label	Number of sentences	Number of characters
ADE[a] suggesting, mean (SD)		
All	3.9 (2.7)	321.7 (456.1)
Criterion A	3.5 (2.6)	399.3 (283.9)
Criterion B	0.4 (0.7)	56.8 (112.3)
Non-ADE suggesting, mean (SD)	48.2 (72.1)	2897.0 (4104.0)

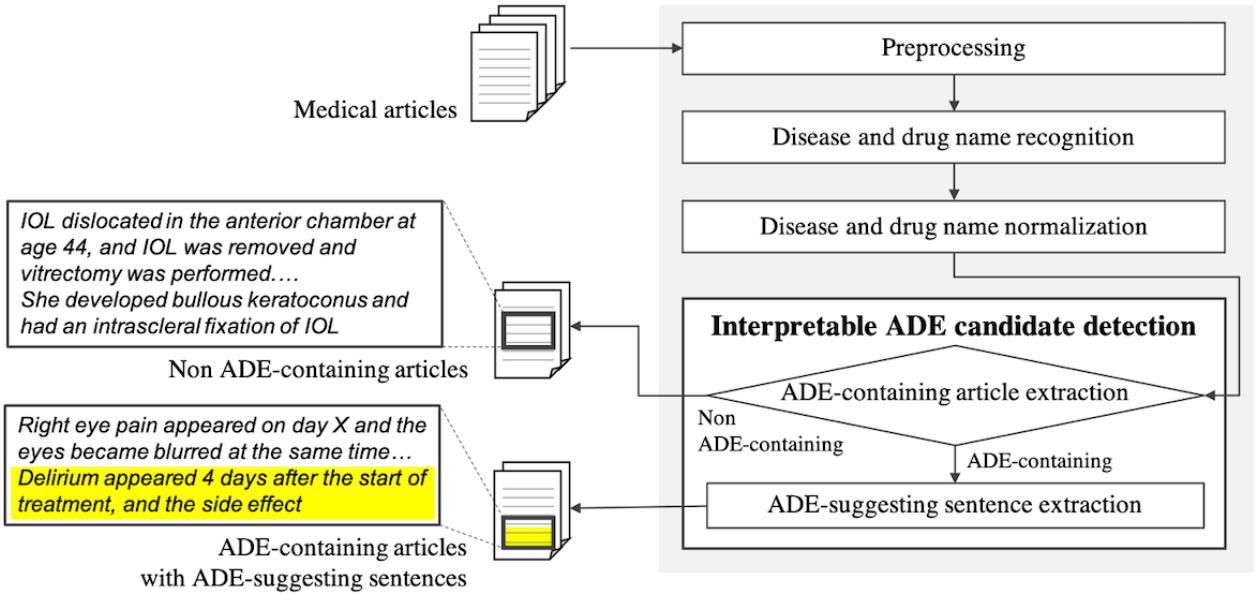
[a]ADE: adverse drug event.

The corpus was annotated by a medical engineer. To evaluate annotation quality, Cohen κ [17] was calculated using parallel annotation data generated by the medical engineer and an author with no prior medical experience. The latter person separately annotated 51/509 (10.0%) of the data set. The Cohen κ values were 0.638 for sentence-level annotation and 0.841 for document-level annotation in the 51/509 (10.0%) sample. Both of these satisfied the standard quality criterion for computational linguistic corpora. Thus, the entire annotation prepared by the medical engineer was adopted for this study.

System Architecture

Figure 1 presents an overview of the architecture of the proposed system. The system comprised preprocessing, disease and drug name recognition, disease and drug name normalization, and interpretable ADE candidate detection submodules.

Figure 1. Architecture of the proposed system. ADE: adverse drug event; IOL: intraocular lens.



1. Preprocessing: Sentences were automatically separated by Japanese full stops and periods only after Japanese letters (including hiragana, katakana, and Chinese characters).
2. Disease and drug name recognition: Disease and drug names were extracted from preprocessed articles. Disease names were extracted with MedEX/J , a disease name extractor provided by Ito et al.[18] which uses conditional random fields (CRFs). This technique is commonly used for named entity recognition and is trained on Japanese case reports. It should be noted that MedEX/J can also extract any English disease names occurring in Japanese medical articles. Drug names were extracted with CRF trained in the same way as MedEX/J. Articles were fed into the models by converting them into the character-based IOB2 format widely used for named entity recognition.
3. Disease and drug name normalization: As there are many variations on the same disease and drug names, they were normalized with the MANBYO [19] and HYAKUYAKU [20] dictionaries. The MANBYO dictionary, the largest Japanese disease name dictionary, can link more than 300,000 disease names (as of September 2020) extracted from Japanese medical documents to various standard forms, such as MedDRA, ICD-10, ICD-11. The HYAKUYAKU dictionary holds more than 40,000 drug names (as of September 2020) extracted from Japanese medical documents and questionnaires to patients and is linked to generic names. These dictionaries also contain English disease/drug names appearing in Japanese medical documents, which can also be normalized.

Distance-based similarity was edited to normalize disease and drug names [21] between extracted and dictionary words. The extracted word was then replaced in the standard expression linked to the dictionary word with the highest similarity. Extracted words were not replaced when they had no dictionary words with similarity exceeding the 0.3

threshold set in this system on the basis of a preliminary experiment.

4. Interpretable ADE candidate detection: This was performed using normalized disease and drug names as features and extracting candidate articles related to ADEs with ADE-suggesting sentences for the second screening.

Interpretable ADE Candidate Detection

Overview

Interpretable ADE candidate detection was conducted to extract useful information for the second screening. ADE-containing article extractions and ADE-suggesting sentence extractions were both performed. Both extractions use binary-classification models. In ADE-containing article extraction, the articles were classified as "ADE containing" or "non-ADE containing." The sentences in "ADE-containing" articles were then classified as "ADE suggesting" or "non-ADE suggesting" in ADE-suggesting sentence extraction. Several design aspects of the system, including the classification algorithm and the feature design used in each model, are described below.

Classification Algorithm

Logistic regression was used to classify the articles and sentences. This method has been widely implemented for text classification. Neural network (NN) models usually outperform other machine learning–based models such as logistic regression in many NLP tasks. However, NN models require much larger corpora and their output is harder to interpret [22]. By contrast, logistic regression can be trained using comparatively less annotated data and the contribution of each feature is easy to determine. Therefore, logistic regression rather NN models was adopted here. The LogisticRegression class of scikit-learn [23] was applied with its default parameters.

Feature Design

An earlier study [24] used the assumption that each sentence refers to at least one disease and drug, and subsequently used identifying features in the words surrounding these key words. Here, however, it was assumed that each sentence does not necessarily refer to any disease or drug. Thus, certain statistical features were created for the setting used here.

For ADE-containing article extraction, expressions alluding to an ADE such as "We stopped the drug" were regarded as important clues for detecting ADEs. The starting point was text in articles as features in orthodox Bag-of-Words representations. MeCab was used to create this Bag-of-Words feature [25]. MeCab is a Japanese morphological analyzer used to separate sentences into words. Those words that appeared only once were removed.

Features concerning diseases and drugs were considered useful for ADE detection as they played key roles in manual ADE detection. Therefore, standard expressions and the sum of their frequency were used as features to account for individual disease and drug characteristics.

To extract ADE-suggesting sentences, the context needs to be considered (as "ADE suggesting" may span multiple sentences). Thus, the features of previous and post sentences in ADE-suggesting sentence extraction, and the same features as ADE-containing article extraction were used. The feature set of interpretable ADE candidate detection is listed below.

- Word tokens: Bag of words appearing in text;
- Standard disease/drug name: Bag of standard disease and drug names;
- Sum of disease/drug name: Sum of occurrence of disease names and sum of occurrence of drug names;
- Context word tokens: Bag of words in previous and post sentences;
- Context standard disease/drug name: Bag of standard disease and drug names in previous and post sentences;
- Context sum of disease/drug name: Sum of occurrence of disease names and sum of occurrence of drug names in previous and post sentences.

The features of each model are shown in Table 2.

Table 2. Feature set used in ADE-containing article extractions and ADE-suggesting sentence extractions.[a]

Feature	ADE[b]-containing article extraction	ADE-suggesting sentence extraction
Word tokens	✓ (7188)	✓ (6597)
Standard disease/drug name	✓ (1043)	✓ (1083)
Sum of disease/drug name	✓ (2)	✓ (2)
Context word tokens	X	✓ (13,194)
Context standard disease/drug name	X	✓ (2166)
Context sum of disease/drug name	X	✓ (4)

[a]The figures in parentheses indicate the average number of variables.

[b]ADE: adverse drug event.

Experiments

Setting

Experiments were conducted to evaluate ADE-containing article extractions and ADE-suggesting sentence extractions.

For ADE-containing article extraction, the classifier trained and predicted the articles by fivefold cross-validation using the features listed in Table 3. All 5 splits of the articles were randomly sampled with the label proportion kept.

Table 3. Effects of each feature in adverse drug event–containing article extraction.

Feature	ΔF1 score
Without word tokens	–0.0456
Without standard disease/drug name	–0.0001
Without sum of disease/drug name	–0.0155

For ADE-suggesting sentence extraction, fivefold cross-validation was applied at the document level. Articles labeled "non-ADE containing" lack sentences labeled "ADE suggesting." Hence, the label proportion was unbalanced when all of the articles in the training set were used for training. To avoid this disequilibrium, only sentences in the "ADE-containing" articles were used for training and all sentences in the test set were used for evaluation.

Evaluation Metrics

Based on the experimental results, F1 scores were calculated to evaluate the performance of our models. To analyze the performance more precisely, we also made precision–recall curves. A precision–recall curve plots recall and precision at each threshold and evaluates tasks with significant trade-offs between measures. High recall or sensitivity means that a model misses no ADEs. This feature is critical for ADE detection. High precision means that the model prediction is reliable. Thus, we must detect ADEs with both reasonable precision and high recall.

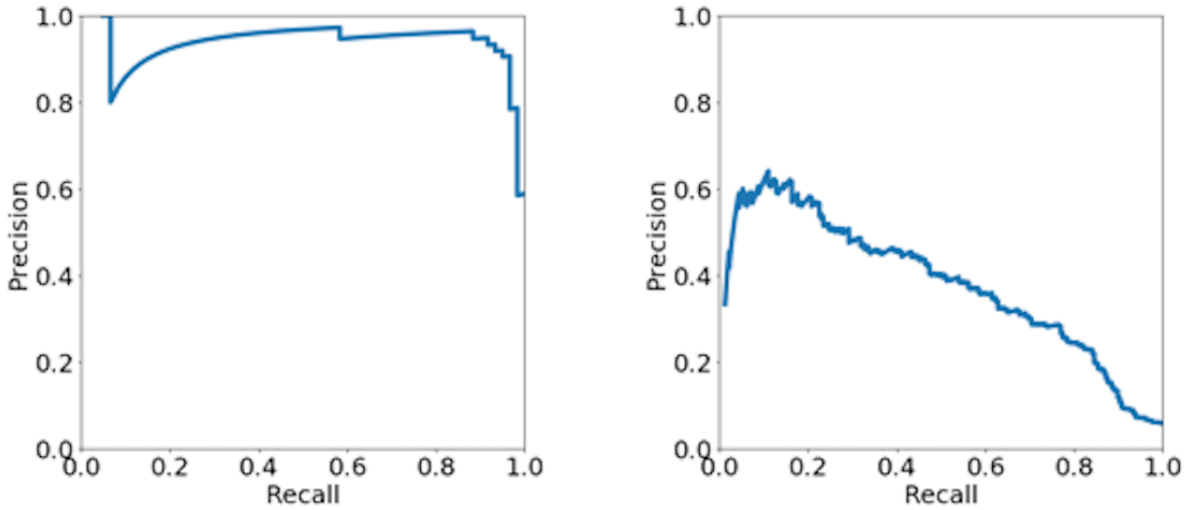
Results

Performance

The average cross-validation for ADE-containing article extraction yielded F1 = 0.903 (SD 0.0165). For ADE-suggesting sentence extraction, F1 was substantially lower; F1 = 0.413 (SD 0.0247).

The precision–recall curves for the validation set with the highest F1 are shown in Figure 2. For ADE-containing article extraction, we achieved high precision and recall. By contrast, both precision and recall were relatively low for ADE-suggesting sentence extraction.

Figure 2. Precision-recall curves for (a) ADE-containing article extraction and (b) ADE-suggesting sentence extraction. ADE: adverse drug event.



Feature Analysis

We conducted an ablation study, which is an evaluation method that removes each feature to quantify the effects of that feature. We removed each feature group and obtained an average F1 score.

Tables 3 and 4 show the results of the ablation studies performed on ADE-containing article extractions and ADE-suggesting sentence extractions in each model. For both models, the Bag-of-Words feature in the corresponding sentence (word tokens) contributed the most. The standard disease and drug names had no influence on classification performance (Table 4). Likewise, the contextual features of standard disease and drug names had no influence on classification performance (Table 4).

Table 4. Effects of each feature in adverse drug event–suggesting sentence extraction.

Feature	ΔF1 score
Without word tokens	–0.0644
Without context word tokens	0.0070
Without standard disease/drug name	0.0
Without context standard disease/drug name	0.0
Without sum of disease/drug name	–0.0204
Without context sum of disease/drug name	–0.0012

Discussion

Feasibility of the Proposed Approach

Performance

The objective of this study was to build a system that supports Japanese drug postmarketing surveillance by automating the first screening and supporting the second screening with medical expertise. Our system effectively addressed this by dividing the task into a relatively easy task, namely, detecting ADEs at the document level, and a comparatively difficult one, that is, detecting ADEs at the sentence level.

Our system classified medical articles related to ADEs with high precision and recall. This result suggests that complex models such as relation classification are unnecessary for this application. Rather, simple document classification suffices to replace manual work in the first screening and thus reduce annotation costs.

The performance of our classification system for ADE-suggesting sentence extraction was relatively poor. However, from the viewpoint of our original goal (supporting experts in drug safety monitoring), performance at this level would still save a large amount of time and cost. Thus, in cases where the model classifies sentences with high recall, there is a high chance of an expert finding the ADE-suggesting sentence after a comparatively short search. In addition, our system was competitive with respect to other relation classification models that extract and classify diseases and drugs according to the relationships among them. The overall performance of ADE-drug relation-based classification is approximately 40%-60% [6].

Feature Contribution

In terms of the feature contribution, word tokens are the features that contributed the most to classify ADE-containing articles and ADE-suggesting sentences. By contrast, the standard disease and drug names and contextual features had less influence on classification performance compared to word tokens. This indicates that the extraction of disease and drug names, which requires relatively large training data to build models, is not necessarily needed to maintain accuracy. All features with their coefficients for ADE-containing article extraction and ADE-suggesting sentence extraction are listed in Multimedia Appendices 2 and 3.

We assumed that language-dependent features such as word embeddings, which are vector representations commonly used to consider the semantics of words, would potentially improve performance. However, obtaining high-quality word embeddings require numerous raw texts, and would be hard to prepare in languages other than English (especially in the medical domain). Thus, we focused more on using language-independent features. The features of each model do not depend on the characteristics of the Japanese language. Therefore, our system is readily applicable to papers written in non-English languages that have relatively small annotated corpora.

Error Analysis

We achieved relatively poor performance for ADE-suggesting sentence extraction. Therefore, we investigated the classification errors of the ADE-suggesting sentence extraction model and used all features for qualitative system output analysis.

Table 5 shows examples of the system classification. For each example, the first sentence is the previous one and the remainder is the corresponding sentence. Note that the terms “gold standard” and “system prediction” represent the corresponding sentence labels. In this section, we analyze cases (c)–(f) that were misclassified by the system.

Table 5. Examples of classification results.

Case	True label	Prediction	Sentences
(a)	ADE[a]	ADE	MTX[b] + adalimumab administration started. Changed to certolizumab pegol because of MTX's side effects.
(b)	ADE	ADE	Case: Male, 74 years old. [Previous history] Rash due to ABPC/SBT[c]. Anaphylactic shock at CEZ[d].
(c)	ADE	Non-ADE	MTX was administered to a 59-year-old patient with RA[e]. In March 201X, she had difficulty breathing and was consulted.
(d)	ADE	Non-ADE	Case: Female, age 79 years. [Chief complaint] [Current medical history] Patient was taking prednisolone 40 mg/d and methotrexate 8 mg/wk. The patient presented with giant cell arteritis and was using insulin for diabetes. Two weeks later, she was hospitalized for malaise and poor appetite.
(e)	Non-ADE[f]	ADE	Stevens–Johnson syndrome (SJS) is characterized by fever and severe mucosal eruptions of the skin, mucosal transitions involving the eyes, lips, and vulva, and blister and erosion due to erythema and necrotic injury to the epidermis. The majority of cases are considered some of the most severe forms of drug eruption. Others are associated with viral and mycoplasmal infections.
(f)	Non-ADE	ADE	Figure 10 shows a clinical course. According to the reporting system, ~25% (92/372) of all drugs causing TdP[g] in the past five years were new quinolones (mainly levofloxacin).
(g)	Non-ADE	Non-ADE	Case: 70-year-old man with a history of hypertension. Right eye pain appeared on day X and was accompanied by blurred vision.
(h)	Non-ADE	Non-ADE	Case: 79-year-old female. The patient had been taking prednisolone 60 mg and methotrexate 6 mg for 6 months following a diagnosis of middle vasculitis.

[a]ADE: ADE suggesting.

[b]MTX: methotrexate.

[c]ABPC/SBT: ampicillin/sulbactam.

[d]CEZ: cefazolin.

[e]RA: rheumatoid arthritis.

[f]Non-ADE: Non-ADE suggesting.

[g]TdP: torsades de pointes.

Cases (c) and (d) are examples wherein the information in the previous sentence was required in order to classify the sentences. Each example would not be regarded as an ADE-suggesting sentence if only the following sentence was considered. However, when the previous sentence was also considered, the following sentence was regarded as ADE suggesting because the symptom mentioned in it may refer to an ADE caused by the drug mentioned in the previous sentence. Classification errors occurred when we added the features of the previous and following sentences.

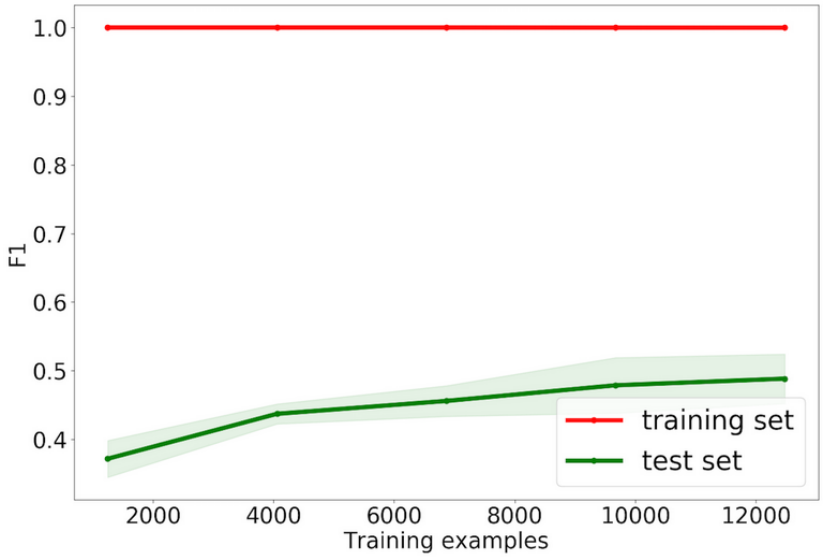
Cases (e) and (f) are examples wherein the general statement is confused with an actual case. The corresponding sentence of each example describes the general disease caused by the drug. However, the general statement and the actual case are similar in terms of their expressions. Consequently, errors resulted.

Limitations

Although our system detects ADE-containing articles with high precision and recall, the performance of ADE-suggesting sentence extraction is relatively poor. There are 3 possible reasons for this poor performance. The first reason is the range of context. Our system can only consider the context for 2 consecutive sentences, and this may increase false negatives. To detect ADEs within a wider context, other approaches would be required such as paragraph classification and sequential labeling, for example, CRF and a hidden Markov model.

The second possible reason is overfitting. Figure 3 shows the F1 scores for several training data sizes. All training set F1 scores were set to 100%. By contrast, those for the validation set were low. This overfitting occurred because the training data size was small relative to the model complexity. In ADE-suggesting sentence extraction, we used 3 times as many features as the ADE-containing article extraction because of the contextual features. Although contextual features contained large amounts of information, most of them were about irrelevant words, which might lead to overfitting.

Figure 3. Training data size versus F1 score in ADE-suggesting sentence extraction. ADE: adverse drug event.



The third possible reason for the poor performance of ADE-suggesting sentence extraction is OCR error. OCR may omit and misread letters, characters, and words, thereby enlarging the vocabulary. It is expected that the improvement of OCR accuracy for Japanese scientific articles consisting of multiple columns will expedite and facilitate preprocessing.

Comparison With Prior Studies

Numerous studies have already identified ADEs reported in medical articles via NLPs [8,9], electronic health records [5,6], and social media posts [10,11]. An annotated corpus for the automatic detection of ADEs was created for case reports in medical articles [26]. Several studies have attempted to detect ADEs by using this corpus [23,27,28]. Various approaches may be used to detect ADEs. However, relation and entity classification were mainly used for this purpose in previous studies. By contrast, our approach is based on document and sentence classification.

- Relation Classification: This approach extracts the relationship between the drug and its corresponding ADE [29,30]. Although it is the most precise way to capture these associations, it entails extensive annotations of all drugs, diseases, and their relationships, which is an expensive process. The classification of drug–disease relationships is difficult when the parameters are only remotely associated [6].
- Entity Classification: This approach focuses unilaterally on ADEs using text written about specific drugs. Diseases are classified only if they are ADEs [11,30]. This approach reduces annotation costs. By contrast, it provides no indication of ADE triggers.
- Sentence Classification: This approach detects ADE-related sentences but does not handle entities. Thus, their relationships to particular drugs are not clarified. The drug and its corresponding ADE appear mainly within a sentence [24]. Nevertheless, if the drug and its corresponding ADEs are separated by more than 1 sentence, this approach would not capture the relationship between them.
- Document Classification: This approach makes ADE-positive or ADE-negative identifications at the document level. In most cases, a document may contain multiple ADEs referred only within that document and all ADEs may be considered simultaneously. However, the output furnishes limited information and manual detection of the ADEs in all sentences is required.

Each of these approaches has both advantages and disadvantages in terms of annotation cost, coverage, and task difficulty. The relation and entity classification methods provide precise information concerning ADEs but their annotation costs are very high. This constraint severely limits their utility for minor languages such as Japanese because comparatively few medical experts are fluent in them. By contrast, document and sentence classification may be conducted at relatively low annotation costs. However, they only detect global phenomena and provide comparatively little information about ADEs. To compensate for the shortcomings of each of these approaches, our system integrated both document and sentence classification.

Conclusions

Here, we developed a system that monitors medical articles for Japanese postmarketing surveillance. Our novel approach, which is based on both document and sentence classification, identifies articles related to ADEs and provides ADE-suggesting sentences. As our system implements a simple classification algorithm, it can be easily applied and managed in-house by pharmaceutical companies.

Our experimental results demonstrate that our system accurately extracts articles related to ADEs. It uses NLP technology which may alleviate some of the manual labor in Japanese pharmaceutical companies.

We aim to apply this system in real-world postmarketing surveillance and evaluate its efficiency and effectiveness in actual monitoring. Going forward, we will explore more complex classification algorithms that can detect a wider range of ADEs.

Acknowledgments

This research is partly supported by Fuji Xerox Co., Ltd. We thank KT and CK for annotating the data set.

Authors' Contributions

SU, SW, SY, and EA designed the study, analyzed the results, and prepared the manuscript. SU implemented the system and conducted the experiments.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Frequency of the drugs appearing in the data.
[XLSX File (Microsoft Excel File), 10 KB - medinform_v8i11e22661_app1.xlsx]

Multimedia Appendix 2
Features and coefficients for ADE-containing article extraction.
[XLSX File (Microsoft Excel File), 252 KB - medinform_v8i11e22661_app2.xlsx]

Multimedia Appendix 3
Features and coefficients for ADE-suggesting sentence extraction.
[XLSX File (Microsoft Excel File), 686 KB - medinform_v8i11e22661_app3.xlsx]

References

1. World Health Organization. International Drug Monitoring: The Role of the Hospital (Report of a WHO Meeting). Geneva, Switzerland: World Health Organization; 1960. URL: https://apps.who.int/iris/bitstream/handle/10665/40747/WHO_TRS_425.pdf?sequence=1&isAllowed=y [accessed 2020-11-17]
2. Howard RL, Avery AJ, Slavenburg S, Royal S, Pipe G, Lucassen P, et al. Which drugs cause preventable admissions to hospital? A systematic review. Br J Clin Pharmacol 2007 Feb;63(2):136-147 [FREE Full text] [doi: 10.1111/j.1365-2125.2006.02698.x] [Medline: 16803468]
3. Rogers AS. Adverse drug events: identification and attribution. Drug Intell Clin Pharm 1987 Nov;21(11):915-920. [Medline: 3678067]
4. Talbot JCC, Nilsson BS. Pharmacovigilance in the pharmaceutical industry. Br J Clin Pharmacol 1998 May 04;45(5):427-431 [FREE Full text] [doi: 10.1046/j.1365-2125.1998.00713.x] [Medline: 9643613]
5. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 01;27(1):3-12. [doi: 10.1093/jamia/ocz166] [Medline: 31584655]
6. Li F, Liu W, Yu H. Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes: Design of an End-to-End Model Based on Deep Learning. JMIR Med Inform 2018 Nov 26;6(4):e12159 [FREE Full text] [doi: 10.2196/12159] [Medline: 30478023]
7. Usui M, Aramaki E, Iwao T, Wakamiya S, Sakamoto T, Mochizuki M. Extraction and Standardization of Patient Complaints from Electronic Medication Histories for Pharmacovigilance: Natural Language Processing Analysis in Japanese. JMIR Med Inform 2018 Sep 27;6(3):e11021 [FREE Full text] [doi: 10.2196/11021] [Medline: 30262450]
8. Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics 2012 Dec 20;3(1):15 [FREE Full text] [doi: 10.1186/2041-1480-3-15] [Medline: 23256479]
9. P Tafti A, Badger J, LaRose E, Shirzadi E, Mahnke A, Mayer J, et al. Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure. JMIR Med Inform 2017 Dec 08;5(4):e51 [FREE Full text] [doi: 10.2196/medinform.9170] [Medline: 29222076]
10. Chen X, Faviez C, Schuck S, Lillo-Le-Louët A, Texier N, Dahamna B, et al. Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate. Front Pharmacol 2018;9:541 [FREE Full text] [doi: 10.3389/fphar.2018.00541] [Medline: 29881351]
11. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-681 [FREE Full text] [doi: 10.1093/jamia/ocu041] [Medline: 25755127]
12. Hans M, Gupta SK. Comparative evaluation of pharmacovigilance regulation of the United States, United Kingdom, Canada, India and the need for global harmonized practices. Perspect Clin Res 2018;9(4):170-174 [FREE Full text] [doi: 10.4103/picr.PICR_89_17] [Medline: 30319947]

13. Food and Drug Administration. Investigational New Drug Safety Reporting Requirements for Human Drug and Biological Products and Safety Reporting Requirements for Bioavailability and Bioequivalence Studies in Humans. Silver Spring, MD: Food and Drug Administration; 2010. URL: https://www.govinfo.gov/content/pkg/FR-2010-09-29/pdf/2010-24296.pdf [accessed 2020-11-17]
14. European Commission. Communication from the Commission – Detailed Guidance on the Collection, Verification and Presentation of Adverse Event/Reaction Reports Arising from Clinical Trials on Medicinal Products for Human Use ('CT-3'). Brussels, Belgium: European Commission; 2011. URL: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2011:172:0001:0013:EN:PDF [accessed 2020-11-17]
15. Pharmaceuticals and Medical Devices Agency. Reports of Side Effects, Infectious Diseases and Defects Based on the Pharmaceutical and Medical Devices Act (for Medical Personnel) (in Japanese). Tokyo, Japan: Pharmaceuticals and Medical Devices Agency URL: https://www.pmda.go.jp/safety/reports/hcp/pmd-act/0003.html [accessed 2020-11-17]
16. NTTDATA NJK Corporation. WinReader PRO v.15.0. URL: https://mediadrive.jp/products/wrp/ [accessed 2020-11-17]
17. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 1960 Apr 01;20(1):37-46. [doi: 10.1177/001316446002000104]
18. Aramaki E, Yano K, Wakamiya S. MedEx/J: A One-Scan Simple and Fast NLP Tool for Japanese Clinical Texts. Stud Health Technol Inform 2017;245:285-288. [Medline: 29295100]
19. Ito K, Nagai H, Okahisa T, Wakamiya S, Iwao T, Aramaki E. J-MeDic: A Japanese disease name dictionary based on real clinical usage. 2018 Presented at: Proceedings of the Eleventh International Conference on Language Resources and Evaluation; May 7–12, 2018; Miyazaki, Japan.
20. Social Computing Laboratory, Nara Institute of Science and Technology. HYAKUYAKU dictionary.. URL: https://sociocom.naist.jp/hyakuyaku-dic-en/ [accessed 2020-11-17]
21. Aramaki E, Imai T, Miyo K, Ohe K. Orthographic disambiguation incorporating transliterated probability. 2008 Presented at: Proceedings of the Third International Joint Conference on Natural Language Processing; Jan 7–12, 2008; Hyderabad, Telangana, India.
22. Belinkov Y, Glass J. Analysis Methods in Neural Language Processing: A Survey. Transactions of the Association for Computational Linguistics 2019 Nov;7:49-72. [doi: 10.1162/tacl_a_00254]
23. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
24. Negi K, Pavuri A, Patel L, Jain C. A novel method for drug-adverse event extraction using machine learning. Informatics in Medicine Unlocked 2019;17:100190. [doi: 10.1016/j.imu.2019.100190]
25. Kudo T. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. URL: https://taku910.github.io/mecab/ [accessed 2020-11-17]
26. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform 2012 Oct;45(5):885-892 [FREE Full text] [doi: 10.1016/j.jbi.2012.04.008] [Medline: 22554702]
27. Kang N, Singh B, Bui C, Afzal Z, van MEM, Kors JA. Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics 2014 Mar 04;15:64 [FREE Full text] [doi: 10.1186/1471-2105-15-64] [Medline: 24593054]
28. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J Biomed Inform 2015 Aug 17 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.013] [Medline: 26291578]
29. Zhao J, Henriksson A, Asker L, Boström H. Predictive modeling of structured electronic health records for adverse drug event detection. BMC Med Inform Decis Mak 2015;15 Suppl 4:S1 [FREE Full text] [doi: 10.1186/1472-6947-15-S4-S1] [Medline: 26606038]
30. Zhao J, Henriksson A, Asker L, Bostrom H. Detecting adverse drug events with multiple representations of clinical measurements. Washington, DC: IEEE; 2015 Presented at: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine; Nov 9–12, 2014; Belfast, UK. [doi: 10.1109/bibm.2014.6999216]

Abbreviations

ADE: adverse drug event
ADR: adverse drug reaction
CRF: conditional random fields
NLP: natural language processing
NN: neural network
OCR: optical character recognition

Edited by G Eysenbach; submitted 21.07.20; peer-reviewed by S Matsuda, W Griffin, A Mahnke; comments to author 12.08.20; revised version received 05.10.20; accepted 28.10.20; published 27.11.20.

Please cite as:
Ujiie S, Yada S, Wakamiya S, Aramaki E
Identification of Adverse Drug Event–Related Japanese Articles: Natural Language Processing Analysis
JMIR Med Inform 2020;8(11):e22661
URL: http://medinform.jmir.org/2020/11/e22661/
doi:10.2196/22661
PMID:33245290

Original Paper

Analysis of Health Insurance Big Data for Early Detection of Disabilities: Algorithm Development and Validation

Seung-Hyun Jeong[1], BS; Tae Rim Lee[1], BS; Jung Bae Kang[2], PhD; Mun-Taek Choi[1], PhD

[1]Sungkyunkwan University, Suwon, Republic of Korea

[2]Korea Disabled People's Development Institute, Seoul, Republic of Korea

Corresponding Author:
Mun-Taek Choi, PhD
Sungkyunkwan University
2066, Seobu-ro, Jangan-gu
Suwon
Republic of Korea
Phone: 82 10 7325 3000
Email: mtchoi@skku.edu

Abstract

Background: Early detection of childhood developmental delays is very important for the treatment of disabilities.

Objective: To investigate the possibility of detecting childhood developmental delays leading to disabilities before clinical registration by analyzing big data from a health insurance database.

Methods: In this study, the data from children, individuals aged up to 13 years (n=2412), from the Sample Cohort 2.0 DB of the Korea National Health Insurance Service were organized by age range. Using 6 categories (having no disability, having a physical disability, having a brain lesion, having a visual impairment, having a hearing impairment, and having other conditions), features were selected in the order of importance with a tree-based model. We used multiple classification algorithms to find the best model for each age range. The earliest age range with clinically significant performance showed the age at which conditions can be detected early.

Results: The disability detection model showed that it was possible to detect disabilities with significant accuracy even at the age of 4 years, about a year earlier than the mean diagnostic age of 4.99 years.

Conclusions: Using big data analysis, we discovered the possibility of detecting disabilities earlier than clinical diagnoses, which would allow us to take appropriate action to prevent disabilities.

(JMIR Med Inform 2020;8(11):e19679) doi:10.2196/19679

KEYWORDS

early detection of disabilities; health insurance; big data; feature selection; classification

Introduction

Providing intervention support early by detecting a child's risk factors for disability helps to prevent not only the disability itself, but also secondary disability by eliminating the risk factors [1-8]. When detection is delayed, the risk of developmental delay is also increased, as the child is unable to perform developmental tasks. If a child's disability is detected after 6 years of age, the child has passed the optimal period of language development, which leads to difficulties in language communication [9].

The main reasons for delayed detection are initial perception by parents, the physician’s wish to delay diagnosis until the prognosis is clearer, or a mistaken assumption by parents that the disorder will improve [3,10]. Childhood developmental delays are difficult to diagnose from a single symptom, as there is a possibility that a temporary delay in development is erroneously considered a disability. If there is no intervention, due to a delay in the detection of the risk of disability in infants and toddlers, the prognosis may not be good [5].

In order to detect a child's disability early, parents must recognize the indications early and request related assistance; policy should provide support to make this possible. However, there is a limit to policy that expands support for assessment costs and reach. Public awareness and education that enables parents to recognize disabilities early should be implemented, but also, in the long run, a system should be established in which the government can identify risk factors in children even if

parents do not recognize them early. It is, therefore, necessary that institutions, such as daycare centers and hospitals, are trained to detect risk factors of disability as soon as possible and to provide parents with relevant information.

Utilizing health insurance big data for early detection may open many possibilities. In South Korea, a system of compulsory medical insurance benefits was initiated in 1977 under the National Health Insurance Act; more than 97% of the public now have obligatory medical insurance, and all related data, including those on diseases and health, are kept and managed by the Korea National Health Insurance Service (KNHIS) [11,12]. With the enactment of the Elderly Long-Term Care Insurance Act in 2007, information relating to the health, nursing, and medical care of older adults is gathered and stored in a cutting-edge information and communications technology database [13]. The data provided by KNHIS contain not only health care provider information but also vast amounts of data (about 2.1 trillion) from people's birth to death [14].

Machine learning techniques that allow computer models to learn knowledge from data [15] can be used to analyze big data such as those in the sample cohort data from KNHIS. Since the medical insurance data contain physician diagnosis records for individuals, the information can be used to label the data, where it becomes a supervised learning problem [16]. Moreover, database classification is a type of supervised learning. It is a process of analyzing existing data to determine the class of newly observed data [17]. Problems that require classification into multiple classes are called multiclass problems.

With the recent availability of national health insurance big data for research purposes, relevant research has commenced. However, since the big data from KNHIS includes sensitive personal information, only some modified data can be used and analyzed through remote access to the KNHIS computer systems. When applying for data export, only deidentified analysis results are made available. Due to these limitations, big data analysis using the health insurance data is still in its infancy [18]. One study [19] that uses KNHIS big data analyzed the correlation between certain diseases, such as sinusitis surgery and asthma. Another study [20] identified diseases that were more likely to occur by using similar group-based data analysis to develop an app service that provides personalized disease and hospital information.

As far as we know, very little research has been done on developing a systematic approach to the early detection of disabilities using big data. Chang [21] examined a supervised learning method for early intervention in children with delayed development based on the clinical data of 516 children below 6 years of age. The study [21] analyzed the association between language, motor, social, and cognitive development from identified diseases, visual problems, psychological and intellectual development, other diseases, and types of delay and, using compositions of the decision tree, made 14 association rules derived scores support and confidence scores. David and Balakrishnan [22] applied a decision tree algorithm and rough sets for the prediction of learning disabilities in school-age children using a checklist of 16 most frequent signs and symptoms of learning disabilities (n=513, area under the receiver operating characteristic curve [AUROC] 0.985). Varol et al [23] present the application of machine learning methods for early prediction of reading disability, collecting 356 samples using 40 features, including demographics, pretesting, and weekly monitoring (word identification fluency); the comparison was made using 6 classification algorithms, and the best result was an AUROC of 0.942. Although these studies [21-23] have showed good learning results on specific disabilities, there are limitations in applying them to all disabilities; since the data used in these studies did not include lifelong records of people with disabilities, temporal tracking for early detection may not be feasible.

The purpose of this study was to detect risk factors for disabilities in children as early as possible based on medical data. Since we conducted early detection analysis on all disabilities, including delayed developmental disabilities, the results are likely to be more meaningful than those of previous studies. By analyzing the effect of each correlation, the disease that is the main cause of the disability could be identified. In this study, various classification algorithms were developed and optimized to find the best model for early detection. As it was based on KNHIS big data, it can lead to more in-depth studies of disabilities in the future.

Our research has the following novelties. As far as we know, it is the first time that a study has investigated early detection using comprehensive disability types using health insurance big data. In order to find the age at which the disability can be diagnosed early, we organized the data by age ranges and created an optimal classification model for each age range. We used multiclass classification algorithms to find the best model for each age range. The earliest age range with clinically significant performance shows the age at which disabilities can be detected early.

Methods

Data

We used medical data extracted from the KNHIS Sample Cohort 2.0 DB, which is an anonymized research database with information on health insurance qualifications, income, history of the hospital and clinic use, and results of health examinations and nursing institutions from 2002 to 2013, covering 1 million people (2% of Korea's 50 million people). Each sample in Sample Cohort 2.0 DB was labeled: no disability, physical disability, brain lesions, visual impairment, hearing impairment, and other disabilities. Other disabilities included all disability types such as speech disability, intellectual disability, and mental disorder. The database contains not only diagnostic codes based on the International Classification of Diseases (ICD) but also additional data such as prescription records, duration of treatments, and frequency of treatments. The distribution of the samples in Sample Cohort 2.0 DB is inherently imbalanced [11,24]. This study complies with the bioethics policy by the institutional review board of Korea National Institute for Bioethics Policy (P01-201905-22-005).

From the raw data, we selected samples for our analysis as follows. The samples we were able to collect at the time of

analysis were records up to the age of 13 years, which would not be an issue for early detection. First, data were extracted from children with acquired disabilities with no missing records from birth to recorded diagnosis, which yielded 804 data records. We selected twice as many data records of children with disabilities, which yielded 1608 data records, to prevent the performance of our analytical model from being distorted by having the number records for those without disabilities being much more than the that of the records with disabilities.

Each sample was identified using a 7-digit personal identification number. Disease diagnostic data and prescription record data were extracted using personal identification numbers. Information on the date of medical treatment and diagnostic codes were available from the disease diagnostic data, classified using disease classification division codes. Prescription record data, such as the date and contents of prescriptions, were extracted from the records. Information on the number of medical actions and prescribed dosage was also recorded.

To discover the age at when the disabilities occurred, the medical records of each sample were organized in units of 1-year increments. The distribution of samples is shown in Table 1. Data for each age range were collected to construct a data set and used for classification learning. In order to improve stability and convergence speed during the optimization process, each feature was transformed to have a mean of 0 and a standard deviation of 1.

Table 1. Data samples by age range.

Age range (years)	No disability	Physical disability	Brain lesions	Visual impairment	Hearing impairment	Other disabilities	Total
Up to 1	1482	40	182	31	47	504	2286
Up to 2	1371	40	182	31	46	504	2174
Up to 3	1263	40	173	30	44	502	2052
Up to 4	1149	40	162	29	41	499	1920
Up to 5	1036	40	147	27	40	489	1779
Up to 6	935	38	137	23	35	473	1641
Up to 7	824	37	122	19	32	446	1480
Up to 8	714	27	102	17	27	400	1287
Up to 9	601	22	84	16	21	324	1068
Up to 10	478	21	70	14	18	265	866
Up to 11	363	18	59	11	15	199	665
Up to 12	242	14	39	6	9	134	444
Up to 13	123	6	17	2	3	71	222

Feature Selection

Feature selection allows selection of a subset of relevant features [25,26]. Good feature selection can make models easier to interpret, shorten learning time, improve learning accuracy, and help avoid the curse of dimensionality [27,28]. We used the extra trees algorithm for feature selection, which is a method of randomly partitioning nodes using a candidate characteristic and then selecting the best partition among them, rather than finding an optimal threshold for partitioning nodes to generate a tree randomly [29]. For the implementation of feature selection, we used ExtraTreeClassifier (scikit-learn, version 0.23.1; Python, version 3.6) [30].

Classification Algorithms

Since there are 6 categories in this study, it is a typical example of multiclass classification. We compared classification algorithms to develop the best model for the early detection of disabilities. We used 4 algorithms in this study: k-nearest neighbor, random forest, logistic regression, and gradient boosting.

The k-nearest neighbor algorithm finds *k* training data closest to the input and uses the output information of these data to estimate the output [31]. Small *k* values indicate a high risk of overfitting, while large values create boundaries with a high propensity to generalization. A variety of methods, such as Euclidean distance, Manhattan distance, and Mahalanobis distance [32], may be used to find adjacent data.

In the random forest model, predictions are generated by bagging several decision trees. Bagging is an ensemble meta-algorithm designed to improve stability and accuracy. Decision trees are similar to the game 20 questions; data are continuously separated based on the characteristics of the data, and the decision tree is classified into 1 correct answer [33,34].

Logistic regression is a linear model that predicts using linear combinations of independent variables [35]. Logistic regression estimates the probability for each group and classifies the data into a group according to a threshold, so it can be applied to the problem of classification [36].

Gradient boosting is a powerful learning algorithm that combines gradient descent with boosting. Gradient descent is an optimization method that reduces error by moving the error

function in the opposite direction to the derivative. Boosting is a method that combines simple and weak learners to make more accurate and powerful learners [37,38]. Even if the accuracy is low, the model compensates for the calculated error [39].

Model Learning

To verify the generalization performance of the model, we divided the data into training data (70%) and test data (30%). Training data were used to train the model; test data were used to evaluate the true classification performance of the trained model.

To find the best model for detecting disabilities, the 4 algorithms were trained. Each classification algorithm has hyperparameters, which when adjusted, show very different performances. Therefore, finding the optimal hyperparameter combination is necessary [30]. We used a grid search to find the optimal combination of hyperparameters for each algorithm. The model was checked against other data to avoid generalization errors during the grid-search process. We used 10-fold cross-validation to avoid further partitioning of data for validation. We used scikit-learn for all implementations.

Performance Metrics

To specify indicators used to evaluate models in this study, we used confusion matrices such as Table 2. The confusion matrix is easy to visually identify when evaluating model performance [40].

Table 2. Confusion matrix for binary classification.

Actual	Predicted	
	Positive	Negative
Positive	True positive	False negative
Negative	False positive	True negative

Accuracy, the most common model performance indicator, is used to show how accurately the model predicts the input data. On the confusion matrix, accuracy is estimated by the sum of the true values divided by the whole; *accuracy* = (*true positive* + *true negative*) / *all*. Precision or the positive predictive value is an indicator of how accurately a model is able to predict a positive; *precision* = *true positive* / (*true positive* + *false positive*). Recall or sensitivity index is the ratio of actual values detected by the model to the actual values; *recall* = *true positive* / (*true positive* + *false negative*). If the data are unevenly distributed, accuracy can lead to distorted performance estimates. The F1 score expresses the harmonic mean of precision and recall. The F1 score gives equal importance to precision and recall. If the data are unevenly distributed, accuracy can lead to distorted performance estimates. Therefore, using F1 scores to measure performance allows for better performance comparisons than those using accuracy [41]; *F1 score* = 2 × *precision* × *recall* / (*precision* + *recall*). The weighted average method was used to measure the average of the indicators for each class; this method assigns a weight according to the number of samples. The weighted average is expressed by the following equation.

where  is the weighted average, x_i is the result from the ith class, N_{class} is the number of classes, and $N_{i\text{-}samples}$ is the number of samples in the ith class.

Results

Early Detection Using Only Disease Diagnostic Data

In our analysis using only ICD disease diagnostic data, we selected the top 150 out of the 4344 disease diagnosis features. Table 3 lists the 10 most important features.

Table 3. Top 10 features in terms of importance when using only disease diagnostic data.

Feature code	Feature name	Importance
F_	Mental and behavioral disorders	0.0498
I10	Essential (primary) hypertension	0.0327
I109	Unspecified hypertension	0.0161
G470	Disorders of initiating and maintaining sleep (insomnias)	0.0145
K259	Unspecified as acute or chronic gastric ulcer without hemorrhage or perforation	0.0133
K590	Constipation	0.0125
E785	Hyperlipidemia, unspecified	0.0120
M4806	Spinal stenosis, lumbar region	0.0120
K295	Chronic gastritis, unspecified	0.0119
J039	Acute tonsillitis, unspecified	0.0114

In model learning, the random forest algorithm performed best across all age ranges (results of the test data set are shown in Table 4). Our aim was to find the earliest age range with an F1 score close to or above 80% to ensure clinical significance [42]. Although the F1 score for up to 6 years was 83.4%, this was not meaningful because the average clinical diagnostic age was 4.99 years according to Sample Cohort 2.0 DB. Up to 4 years had an F1 score of 79.6%, which is close to 80%, and the age range is clinically meaningful. This model would detect disability almost 1 year earlier, given that the average clinical diagnostic age is 4.99 years.

Table 4. Model learning results when using only disease diagnostic data.

Age range (years)	Classifier	Parameters	Accuracy	Precision	Recall	F1 score
Up to 1	Random forest	n estimators: 16	0.703	0.639	0.703	0.660
Up to 2	Random forest	n estimators: 64	0.758	0.718	0.758	0.725
Up to 3	Random forest	n estimators: 64	0.800	0.778	0.800	0.776
Up to 4	Random forest	n estimators: 64	0.816	0.798	0.816	0.796
Up to 5	Random forest	n estimators: 64	0.818	0.787	0.818	0.796
Up to 6	Random forest	n estimators: 128	0.852	0.833	0.852	0.834
Up to 7	Random forest	n estimators: 64	0.836	0.805	0.836	0.813
Up to 8	Random forest	n estimators: 64	0.850	0.836	0.850	0.835
Up to 9	Random forest	n estimators: 128	0.854	0.837	0.854	0.838
Up to 10	Random forest	n estimators: 128	0.852	0.832	0.852	0.836
Up to 11	Random forest	n estimators: 64	0.873	0.854	0.873	0.856
Up to 12	Random forest	n estimators: 128	0.864	0.866	0.864	0.863
Up to 13	Random forest	n estimators: 64	0.922	0.929	0.922	0.914

The confusion matrix of the analysis for the range up to 4 years is given in Table 5. As the model was learned, the average for each class was high. Thus, the results of the confusion matrix indicate that most samples for children without disabilities were well classified.

Table 5. Confusion matrix when using only disease diagnostic data.

Actual	Predicted					
	No disability	Physical disability	Brain lesions	Visual impairment	Hearing impairment	Other disabilities
No disability	334	0	0	0	0	11
Physical disability	7	0	0	0	0	5
Brain lesions	4	0	34	0	0	10
Visual impairment	6	0	1	1	0	1
Hearing impairment	1	0	0	0	6	5
Other disabilities	46	0	9	0	0	95

Early Detection Using Disease Diagnostic and Prescription Data

A second analysis was performed by adding prescription record data to the disease diagnostic data used in the previous analysis. Prescription data included information on medications, treatment materials, and medical practices received by patients. We used the top 150 out of a total of 12,713 features, including 4344 diseases and 8369 prescription data. Table 6 lists the 10 most important features.

Table 6. Top 10 features in terms of importance when using disease diagnostic and prescription data

Feature code	Feature name	Importance
F6203	Social Maturity Scale	0.0215
F_	Mental and behavioral disabilities	0.0124
F6201	Intelligence test	0.0123
NN011	Personal supportive psychotherapy	0.0105
F6215	Personality test (pictorial test)	0.0089
FY731	Childhood Autism Rating Scale	0.0087
NN031	Family therapy	0.0075
130801ASY	Hypnotic sedatives	0.0063
NN013	Personal intensive psychotherapy	0.0060
F6240	Bender Gestalt Test	0.0057

In model learning, both random forest and gradient boosting algorithms performed well (Table 7). In this analysis, the F1 score of the up to 4-year age range was 81.6%, which indicates that the early detection of disabilities seems to be relatively certain. In addition, as the F1 score for the up to 3-year age range was 78.3%, it is possible that improvements could lead to a diagnosis about 2 years before 4.99 years.

Table 7. Model learning results based on disease diagnostic and prescription data.

Age range (years)	Classifier	Parameters	Accuracy	Precision	Recall	F1 score
Up to 1	Logistic regression	*C*=0.1	0.732	0.691	0.732	0.688
Up to 2	Gradient boosting	learning rate: 0.4; n estimators: 4	0.767	0.743	0.767	0.738
Up to 3	Random forest	n estimators: 128	0.802	0.800	0.802	0.783
Up to 4	Random forest	n estimators: 128	0.832	0.819	0.832	0.816
Up to 5	Random forest	n estimators: 32	0.835	0.813	0.835	0.817
Up to 6	Gradient boosting	learning rate: 0.4; n estimators: 4	0.858	0.850	0.858	0.853
Up to 7	Random forest	n estimators: 32	0.849	0.830	0.849	0.834
Up to 8	Random forest	n estimators: 128	0.866	0.848	0.866	0.854
Up to 9	Gradient boosting	learning rate: 0.4; n estimators: 4	0.857	0.859	0.857	0.857
Up to 10	Random forest	n estimators: 128	0.898	0.878	0.898	0.885
Up to 11	Random forest	n estimators: 64	0.914	0.916	0.914	0.905
Up to 12	Gradient boosting	learning rate: 0.4; n estimators: 1	0.832	0.833	0.832	0.829
Up to 13	Gradient boosting	learning rate: 1.0; n estimators: 1	0.891	0.896	0.891	0.893

The confusion matrix of the analysis for the range up to 4 years is given in Table 8. As this was a learned model, the average for each class was high. The results of the confusion matrix, therefore, indicate that most children without disabilities were correctly classified. Children with physical disabilities were still not well classified, but there was some improvement in most classes.

Table 8. Confusion matrix when using disease diagnostic and prescription data.

Actual	Predicted					
	No disability	Physical disability	Brain lesions	Visual impairment	Hearing impairment	Other disabilities
No disability	336	0	0	0	0	9
Physical disability	6	0	5	0	0	1
Brain lesions	4	0	36	0	0	11
Visual impairment	4	0	1	4	0	1
Hearing impairment	3	0	0	0	5	4
Other disabilities	35	0	17	0	0	98

Discussion

In this study, we used big data analysis for early detection of children who are more likely to have disabilities. An analysis of the sample data suggests that it is possible to detect disability early with accuracy at 3 or 4 years, which is before the average diagnostic age of 4.99 years. This means that children who may be at risk of disability due to various risk factors can be screened early based on medical records alone and can receive appropriate treatment to reduce the degree of disability.

The contributions of our study are described as follows. Our study is one of the first to investigate early detection of disabilities, covering all disabilities comprehensively based on KNHIS big data. This shows that health insurance data is of great value in analyzing disabilities and provides a basis for future studies. To find the age at which disabilities can be detected early, we set up a multiclass classification frame that organizes data by age ranges and trains multiple algorithms to select the best model. This frame can be further improved so that it could be an important tool for experts in the field.

Our study has the following limitations. Though it would be better if the disability was detected by age 3 years or earlier, the early detection performance from the up to 3-year age range did not exceed the clinically significant threshold of 80% due to limitations in health insurance sample data. Another limitation was that the other category of disabilities hampered the performance of the model. Future research with more data and detailed classification of other types of disabilities could lead to a more accurate analysis. The imbalance of samples also had an important impact on data analysis. In this analysis, the number of children with disabilities was 804; of which, 504 had other types of disabilities. Since data on physical disability, visual impairment, and hearing impairment were relatively less, the model may not have learned sufficiently; therefore, it is necessary to ensure that there is sufficient data for each type when conducting further studies. We chose the best model based on the F1 score, but in practice, depending on the situation, we may choose the best model with the least false positives or false negatives.

To improve the early detection model in the future, the following work can be done in the future. In addition to the records of diagnosed diseases and prescription medications used in this analysis, various data such as health medical examination data, are also collected by the National Health Insurance Service. Incorporating these additional data to overcome the abovementioned limitations could lead to the development of more sophisticated models for early disability detection analysis. Moreover, feature engineering is important because the number of features can increase tremendously, and future studies require a more diverse application and comparison of feature engineering algorithms.

Acknowledgments

This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (20003762) funded by the Korean Ministry of Trade, Industry, & Energy and Korea Disabled People's Development Institute Institutional Program (grant number: Policy19-18).

Conflicts of Interest

None declared.

References

1. Kim JH, Cha JK. [Analysis of research studies about at-risk children]. Journal of Emotional & Behavioural Disorders 2015;31(3):127-151 [FREE Full text]
2. Dawson G, Watling R. Interventions to facilitate auditory, visual, and motor integration in autism: a review of the evidence. J Autism Dev Disord 2000 Oct;30(5):415-421. [doi: 10.1023/a:1005547422749] [Medline: 11098877]
3. Lee KS, Jung SJ, Park JA, Shin YJ, Yoo HJ. Factors of early screening of young children with autism spectrum disorder. Journal of the Korean Association For Persons With Autism 2015;15(3):1-24 [FREE Full text]

4. Lee SH, Lee SJ. Trends and issues in research regarding young children with autism spectrum disorders in Korea. The Korean Journal of Early Childhood Special Education 2012 Jun 30;12(2):23-53 [FREE Full text]
5. Lee J, Kim Y, Hwang Y, Ko J. A Study on the Effect and Administrative Support of Early Intervention for Young Children-At-Risk. Korea Assoc Early Child Educ Educ Welf 2018 Sep 30;22(3):173-209. [doi: 10.22590/ecee.2018.22.3.173]
6. Lee HS. A Study on Development of Checklist for Autistic Disorder in Infant and Toddler. Journal of Special Education & Rehabilitation Science 2008 Dec;47(4):65-90 [FREE Full text]
7. McDonnell AP. Dealing with individual differences in the early childhood classroom. Journal of Early Intervention 2016 Sep 14;19(1):87-90. [doi: 10.1177/105381519501900109]
8. Song M, Choi Y. Special child consultation [Korean]. Seoul: SigmaPress; Mar 05, 2013.
9. Chakrabarti S, Fombonne E. Pervasive developmental disorders in preschool children. JAMA 2001 Jun 27;285(24):3093-3099. [doi: 10.1001/jama.285.24.3093] [Medline: 11427137]
10. Kim KH. Mothers? expectations of young children with autism in education. Journal of Special Education: Theory and Practice 2014 Dec;15(4):535-558 [FREE Full text]
11. Sample cohort 2.0 DB user manual ver1.0. National Health Insurance Service. 2017 Jun. URL: http://medical.yonsei.ac.kr/we/?module=file&act=procFileDownload&file_srl=459938&sid=4ba5676b6fc2b61886ccf788504e2056&module_srl=584 [accessed 2019-08-15]
12. Lee J, Lee JS, Park S, Shin SA, Kim K. Cohort Profile: The National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. Int J Epidemiol 2017 Apr 01;46(2):e15. [doi: 10.1093/ije/dyv319] [Medline: 26822938]
13. Kwon S. Thirty years of national health insurance in South Korea: lessons for achieving universal health care coverage. Health Policy Plan 2009 Jan;24(1):63-71. [doi: 10.1093/heapol/czn037] [Medline: 19004861]
14. Cho S, Kim H, Kang G. A visual query database system for the Sample Research DB of the National Health Insurance Service. Korean Journal of Applied Statistics 2017 Feb 28;30(1):13-24. [doi: 10.5351/kjas.2017.30.1.013]
15. Michie D, Spiegelhalter D, Taylor C. Machine learning, Neural, and Statistical Classification. United States: Ellis Horwood; 1994.
16. Kotsiantis S, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Emerging artificial intelligence applications in computer engineering 2007;160(1):3-24.
17. Aly M. Survey on multiclass classification methods. 2005. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.175.107&rep=rep1&type=pdf [accessed 2020-11-14]
18. Jang JS, Cho SH. Mobile Health (m-health) on Mental Health. Korean J Str Res 2016 Dec 31;24(4):231-236. [doi: 10.17547/kjsr.2016.24.4.231]
19. Yu S, Wee J, Kim J, Yoon S. Methodology for Big Data Analysis Using Data from National Health Insurance Service: Preliminary Methodologic Study and Review about the Relationship between Sinus Surgery and Asthma. J Rhinol 2015;22(1):28. [doi: 10.18787/jr.2015.22.1.28]
20. Kim SH, Hwang HS. Developing a Personalized Disease and Hospital Information Application Using Medical Big Data. Entrue Journal of Information Technology 2016;15(2):7-16 https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002174587.
21. Chang C. A study of applying data mining to early intervention for developmentally-delayed children. Expert Systems with Applications 2007 Aug;33(2):407-412. [doi: 10.1016/j.eswa.2006.05.007]
22. David JM, Balakrishnan K. Machine Learning Approach for Prediction of Learning Disabilities in School-Age Children. IJCA 2010 Nov 10;9(11):7-14. [doi: 10.5120/1432-1931]
23. Varol HA, Mani S, Compton DL, Fuchs LS, Fuchs D. Early prediction of reading disability using machine learning. AMIA Annu Symp Proc 2009 Nov 14;2009:667-671 [FREE Full text] [Medline: 20351938]
24. Number of Registered Persons with Disabilities and Disability Pension Recipients. Ministry of Health and Welfare of South Korea. URL: http://www.mohw.go.kr/eng/hs/hs0106.jsp?PAR_MENU_ID=1006&MENU_ID=100606 [accessed 2019-08-15]
25. Guyon I, Gunn S. Feature Extraction: Foundations and Applications. Heidelberg: Springer; 2006.
26. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018 Jul;300:70-79. [doi: 10.1016/j.neucom.2017.11.077]
27. Kim Y, Kwon K. Improvement of Classification Accuracy on Success and Failure Factors in Software Reuse using Feature Selection. KIPS Transactions on Software and Data Engineering 2013 Apr 30;2(4):219-226. [doi: 10.3745/ktsde.2013.2.4.219]
28. DASH M, LIU H. Feature selection for classification. Intelligent Data Analysis 1997;1(1-4):131-156. [doi: 10.1016/s1088-467x(97)00008-5]
29. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006 Mar 2;63(1):3-42. [doi: 10.1007/s10994-006-6226-1]
30. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. GetMobile: Mobile Comp. and Comm 2015 Jun;19(1):29-33. [doi: 10.1145/2786984.2786995]
31. Peterson L. K-nearest neighbor. Scholarpedia 2009;4(2):1883. [doi: 10.4249/scholarpedia.1883]
32. Weinberger K, Saul L. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research 2009;10(2):207-244.
33. Liaw A, Wiener M. Classification and regression by RandomForest. R news 2002;2(3):18-22.

34. Zhang C, Ma Y. Ensemble Machine Learning. Boston: Springer; 2012.
35. Kleinbaum D, Dietz K. Logistic regression. New York: Springer-Verlag; 2002:9781441917423.
36. Kleinman LC, Norton EC. What's the Risk? A simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. Health Serv Res 2009 Feb;44(1):288-302 [FREE Full text] [doi: 10.1111/j.1475-6773.2008.00900.x] [Medline: 18793213]
37. Bottou L. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010.: Physica-Verlag HD; 2010 Presented at: 19th International Conference on Computational Statistics; August 22-27; Paris France p. 177-186. [doi: 10.1007/978-3-7908-2604-3_16]
38. Schapire RE, Freund Y. Boosting: Foundations and algorithms. Cambridge, Massachusetts: MIT Press; 2013.
39. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21 [FREE Full text] [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
40. Powers DM. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Journal of Machine Learning Technologies 2011;2(1):37-63 http://www.bioinfo.in/contents.php?id=51 / The journal ceased publication and is no longer accepting submissions (https://bioinfopublication.org/pages/journal.php?id=BPJ0000274) [FREE Full text]
41. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: AI 2006: Advances in Artificial Intelligence.: Springer; 2006 Presented at: Australasian Joint Conference on Artificial Intelligence; 2006; Berlin p. a. [doi: 10.1007/11941439_114]
42. Nunnally JC. Psychometric Theory— 25 Years Ago and Now. Educational Researcher 2016 Jul;4(10):7-21. [doi: 10.3102/0013189x004010007]

Abbreviations

AUROC: area under the receiver operating characteristic curve
ICD: International Classification of Diseases
KNHIS: the Korea National Health Insurance Service

Edited by G Eysenbach; submitted 30.04.20; peer-reviewed by KL Ong, V Bremer; comments to author 02.07.20; revised version received 27.07.20; accepted 30.10.20; published 23.11.20.

Please cite as:
Jeong SH, Lee TR, Kang JB, Choi MT
Analysis of Health Insurance Big Data for Early Detection of Disabilities: Algorithm Development and Validation
JMIR Med Inform 2020;8(11):e19679
URL: http://medinform.jmir.org/2020/11/e19679/
doi:10.2196/19679
PMID:33226352

Original Paper

Alert Override Patterns With a Medication Clinical Decision Support System in an Academic Emergency Department: Retrospective Descriptive Study

Junsang Yoo[1], PhD; Jeonghoon Lee[2], MD; Poong-Lyul Rhee[3], MD, PhD; Dong Kyung Chang[2,3,4], MD, PhD; Mira Kang[2,4,5], MD, PhD; Jong Soo Choi[4], PhD; David W Bates[6], MD, MCs; Won Chul Cha[2,4,7], MD

[1]Institution of Healthcare Resource, School of Nursing, Sahmyook University, Seoul, Republic of Korea

[2]Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Department of Digital Health, Sungkyunkwan University, Seoul, Republic of Korea

[3]Department of Gastroenterology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[4]Health Information and Strategy Center, Samsung Medical Center, Seoul, Republic of Korea

[5]Center for Health Promotion, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[6]Division of General Internal Meidicine and Primary Care, Brigham and Women's Hospital, Boston, MA, United States

[7]Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

Corresponding Author:
Won Chul Cha, MD
Department of Emergency Medicine
Samsung Medical Center
Sungkyunkwan University School of Medicine
81 Irwon-ro
Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 10 5386 6597
Email: docchaster@gmail.com

Abstract

Background: Physicians' alert overriding behavior is considered to be the most important factor leading to failure of computerized provider order entry (CPOE) combined with a clinical decision support system (CDSS) in achieving its potential adverse drug events prevention effect. Previous studies on this subject have focused on specific diseases or alert types for well-defined targets and particular settings. The emergency department is an optimal environment to examine physicians' alert overriding behaviors from a broad perspective because patients have a wider range of severity, and many receive interdisciplinary care in this environment. However, less than one-tenth of related studies have targeted this physician behavior in an emergency department setting.

Objective: The aim of this study was to describe alert override patterns with a commercial medication CDSS in an academic emergency department.

Methods: This study was conducted at a tertiary urban academic hospital in the emergency department with an annual census of 80,000 visits. We analyzed data on the patients who visited the emergency department for 18 months and the medical staff who treated them, including the prescription and CPOE alert log. We also performed descriptive analysis and logistic regression for assessing the risk factors for alert overrides.

Results: During the study period, 611 physicians cared for 71,546 patients with 101,186 visits. The emergency department physicians encountered 13.75 alerts during every 100 orders entered. Of the total 102,887 alerts, almost two-thirds (65,616, 63.77%) were overridden. Univariate and multivariate logistic regression analyses identified 21 statistically significant risk factors for emergency department physicians' alert override behavior.

Conclusions: In this retrospective study, we described the alert override patterns with a medication CDSS in an academic emergency department. We found relatively low overrides and assessed their contributing factors, including physicians' designation and specialty, patients' severity and chief complaints, and alert and medication type.

(JMIR Med Inform 2020;8(11):e23351) doi:10.2196/23351

KEYWORDS

medical order entry systems; decision support systems; clinical; alert fatigue; health personnel; clinical decision support system; alert; emergency department; medication

Introduction

An emergency department (ED) is a challenging environment in which multiple interventions are delivered within a short period [1]. The severity of the patients' conditions demands that providers often order medications and tests simultaneously, which could contribute to a higher rate of medical errors [2-4]. Physicians working in an ED must often make decisions in the context of uncertainty due to the pace of the environment and resource limitations [5]. Specifically, the concept of physicians working in an ED is not limited to emergency medicine specialists, but rather covers various medical department physicians who treat patients in the geographical area of the ED.

Computerized provider order entry (CPOE) combined with a clinical decision support system (CDSS) was introduced to reduce preventable adverse drug events [6]. This system was expected to improve physicians' prescribing patterns by supporting their decision-making process in a variety of ways. However, previous studies have revealed that physicians' override rates on CDSS alerts are high [7-10], raising concerns about the effectiveness of CDSSs in many implementations [11-13].

Many factors, including physician and patient characteristics, environmental factors, and factors associated with the system itself, affect physicians' alert override patterns in multifactorial ways with probable interactions among them [14-16]. Additionally, many previous studies regarding physicians' alert override patterns have focused on specific diseases or alert types for well-defined targets as well as particular settings [10,17,18]. Thus, it is not clear how these results will generalize to patients at large or in settings such as the ED. Moreover, alert-related fatigue and physician burnout are very frequent among ED physicians, and also appear to be associated with worse performance of a CDSS [19-21].

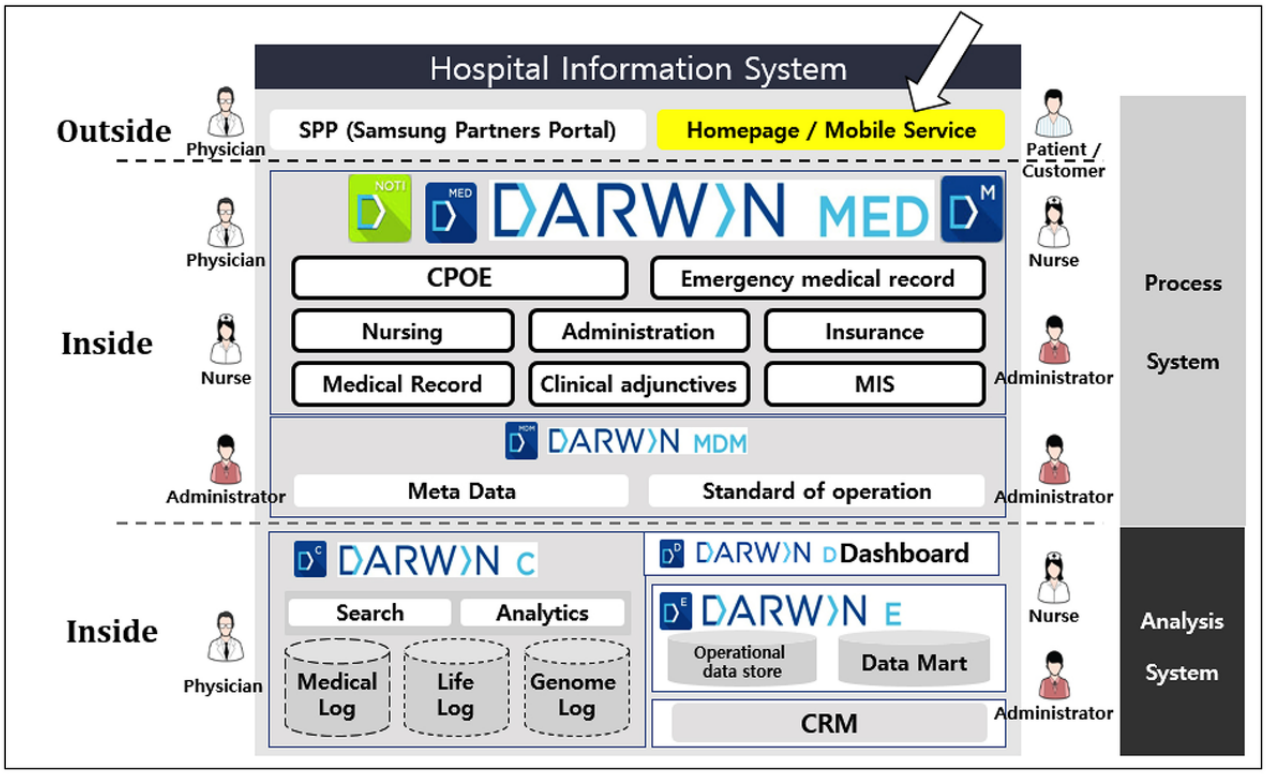
Based on this background, the aim of this study was to describe and assess alert override patterns with a medication CDSS in a large academic ED.

Methods

Study Setting

This study was conducted at an ED with an annual visit volume of 80,000 patients. The hospital is an academic institute with 2000 inpatient beds. The institution has utilized a home-grown electronic health record (EHR) system since 2003, which was replaced by a next-generation EHR system named Data Analytics and Research Window for Integrated Knowledge (DARWIN) in 2016. DARWIN is an all-in-one home-grown EHR that includes CPOE, nursing, pharmacy, billing, research support, and a patient portal (Figure 1) [22]. The institution's ethics committee approved this study (Institutional Review Board File No. 2019-05-038).

Figure 1. Overall schematic description of the hospital information system architecture at the Samsung Medical Center. DARWIN: Data Analytics and Research Window for Integrated Knowledge; CPOE: computerized physician order entry; MIS: management information system; MDM: master data management; CRM: customer relationship management. Reproduced with permission from Jung et al [22].



Minimally Interruptive CDSS

When developing DARWIN's CDSS, a minimally interruptive medication CDSS was introduced. This CDSS is mainly designed for physicians and utilizes only medication-specific information so that, for instance, there is no interference with laboratory data. This database is supplied from Medi-Span (Wolters Kluwer Health, Philadelphia, PA, USA) and is updated monthly (Figure 2).

The user interface was designed to minimize interruption in physician prescription workflow. First, the rules engine operates simultaneously with the physician's entry of each order component such as a drug name, dose, and route. Second, its feedback appears as an in-line message so that physicians are not interrupted during order processing (Figure 3). The CDSS operates with the following areas of medication: age, allergy, disease, duplication, gender, lactation, pregnancy, route, drug-drug interaction, and dosage.

Figure 2. System architecture of the computerized provider order entry (CPOE). DB: database.

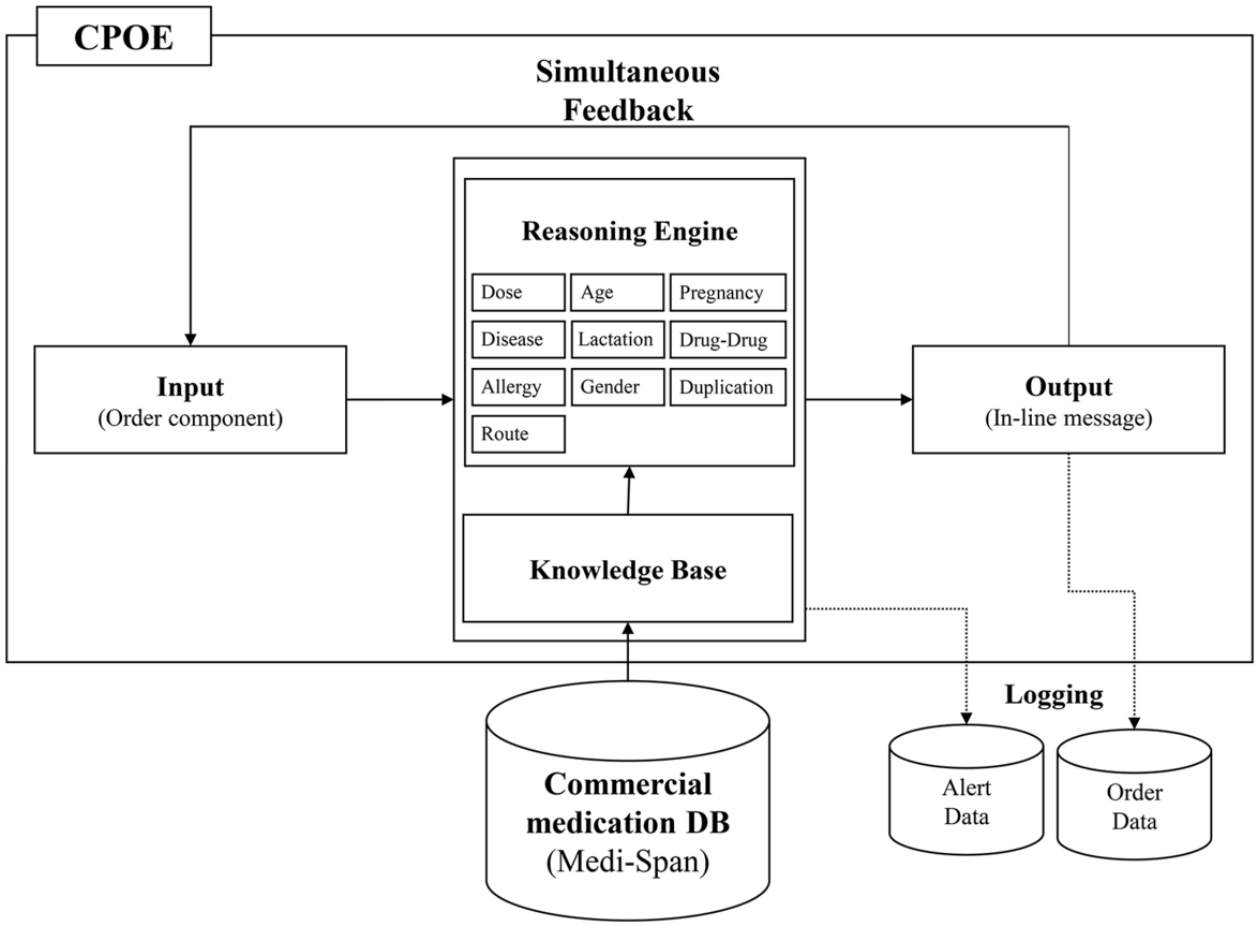
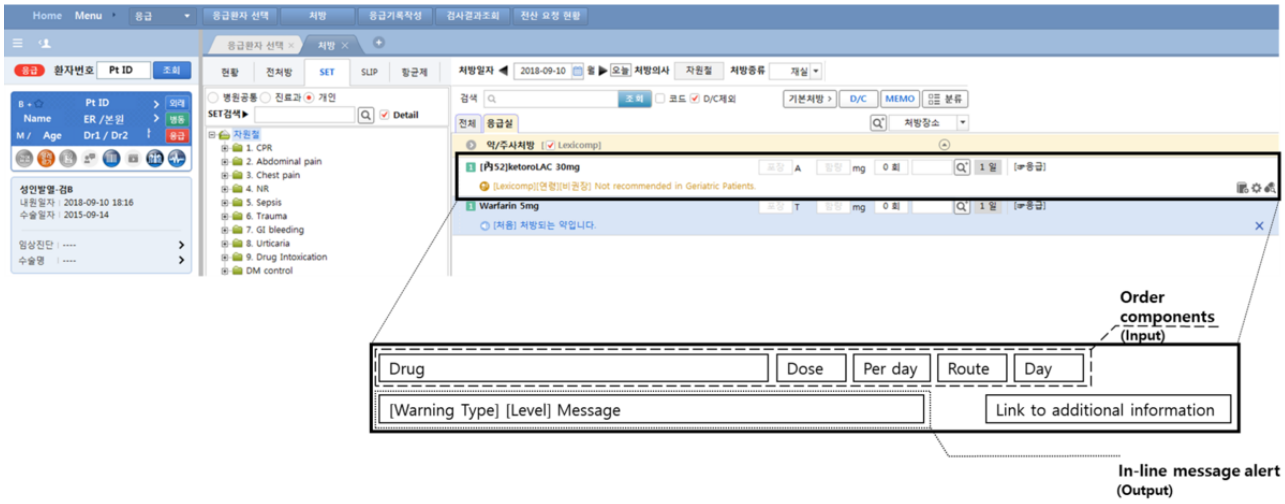


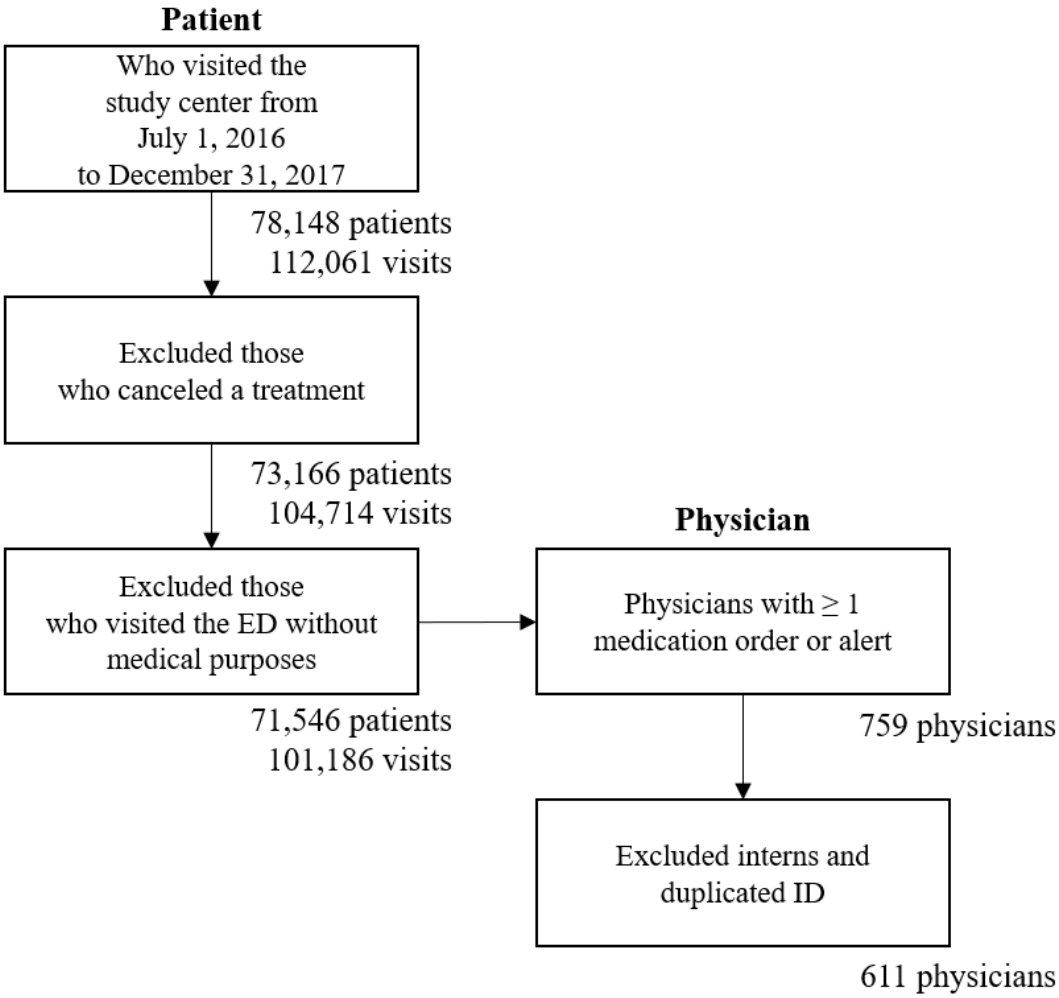
Figure 3. Screenshot of the computerized provider order entry system and features of its interface (zoomed out).



Study Subjects

The inclusion criteria for this study were that patients had to have visited the ED between July 1, 2016 and December 31, 2017. Patients were excluded if they visited the ED but left without being examined by physicians or if they visited the ED without a medical purpose. The eligibility and selection process is presented in Figure 4. As we aimed to extensively investigate alert override patterns in an ED, the term “physician in ED” includes physicians from various medical departments, including the ED, pediatrics, internal medicine, and plastic surgery, among others.

Figure 4. Flow diagram of the eligibility and selection process for study inclusion. ED: emergency department; ID: identification.



Data Extraction and Preparation

Clinical data were extracted from the clinical data warehouse of the study site. We collected the following data: patient information (deidentified patient identifier, date of birth, gender, chief complaint, visit time, type of disposition, length of ED stay, severity level, and International Classification of Diseases-10 code), alert information (medication code based on the generic product identifier, alert firing time), order information (medication code, order time), and physician information (physician identifier, department, and career status). The severity score was measured by the Korean Triage and Acuity Scale (KTAS), which has been widely used by triage nurses in Korean EDs [23]. The KTAS was developed based on the Canadian Triage Scale to assess ED visiting patients' acuity and severity. Patients with a score corresponding to level 1 have the highest acuity and severity, whereas level 5 indicates the lowest acuity and severity.

Override Determining Algorithm

An override determining algorithm was developed for assessing the outcome measures. The algorithm was based on a rule-based, alert type–specific logic, newly generated for this study for validation (Table 1).

Table 1. Description of the alert overrides determining logic.

Type of alert	Override determining logic
Age, allergy, disease, duplication, gender, lactation, pregnancy, route	If a physician completed the order without alert adjustment
Dose	If a physician did not adjust the dose-related order components such as prescription day or daily dosage
Drug-drug interaction	If a physician ordered both medications indicated in a drug-drug interaction alert

For the chart review, we selected a sample of 20 alerts among each type of alert that was performed for both overridden and nonoverridden orders. In the first round of the review process, two clinicians independently reviewed the sample alerts and then evaluated the interrater reliability using the Cohen κ statistic. In the second round, both clinicians worked together

to resolve any case of disagreement. The two clinicians who reviewed the log data consisted of a doctor and nurse who have worked at the ED of the study site for over 4 years. Accuracy of the override determining algorithm was assessed using the reviewed data as the gold standard.

Data Analysis

We conducted a descriptive analysis of patients, physicians, alert characteristics, and the alert firing and override rates. We used a logistic regression model for assessing the risk of the alert override using scalable medical variables such as physician factors (physicians' specialties and designation), patient factors (severity scores and chief complaints), and alert factors (types of alerts and medication categories of alerts). The statistical significance level was set at $P<.05$. The variable with the smallest difference between the overall mean override rate and overrode rate of each variable (eg, resident) within each group (eg, physicians' designation) was selected as the reference variable of the logistic regression. We employed R (version 3.6.0) software for the analysis.

Results

Interrater Reliability of the Override Determining Algorithm

In the first round of the review process conducted by two independent clinicians, Cohen κ was 0.82 (95% CI 0.74-0.90). All discrepancies were resolved in the second round of the review process. The accuracy of the override determining algorithm was 0.95.

Basic Characteristics

During the study period, 611 physicians took care of 71,546 patients with 101,186 visits. General characteristics of the physicians and patients are described in Table 2 and Table 3, respectively. The physicians prescribed 748,339 medication orders and 102,887 (13.75%) alerts were fired.

Table 2. Physician characteristics (N=611).

Characteristic	n (%)
Designation	
Resident	357 (58.3)
Fellow	154 (25.2)
Faculty	100 (16.4)
Specialty	
Emergency Medicine	41 (6.7)
General Internal Medicine	60 (9.8)
Gastroenterology	39 (6.4)
Cardiology	17 (2.8)
Pulmonary Medicine	15 (2.5)
Nephrology	13 (2.1)
Hematology & Oncology	10 (1.6)
Endocrinology & Metabolism	8 (1.3)
Infectious Disease	7 (1.2)
Allergic Medicine	2 (0.3)
Rheumatology	2 (0.3)
General Surgery	58 (9.5)
Gynecology & Obstetrics	39 (6.4)
Thoracic surgery	26 (4.3)
Orthopedic Surgery	19 (3.1)
Neurosurgery	15 (2.5)
Urology	15 (2.5)
Plastic Surgery	11 (1.8)
Pediatrics	53 (8.7)
Family Medicine	24 (3.9)
Ophthalmology	24 (3.9)
Otolaryngology	24 (3.9)
Neurology	20 (3.3)
Radiology	16 (2.6)
Psychiatry	14 (2.3)
Anesthesiology & Pain Medicine	10 (1.6)
Dermatology	9 (1.5)
Critical Care Medicine	7 (1.2)
Rehabilitation	7 (1.2)
Dentistry	5 (0.8)
Radiation Oncology	1 (0.2)

Table 3. Patient characteristics (N=101,186 visits).

Characteristic	Value
Age (years), mean (SD)	44.50 (25.68)
Male, n (%)	51,221 (50.62)
Severity score, n (%)	
1 (Highest severity)	1122 (1.11)
2	6331 (6.25)
3	38,456 (38.01)
4	46,696 (46.15)
5 (Lowest severity)	8581 (8.48)
Chief complaint, n (%)	
Fever	15,080 (14.90)
Abdominal Pain	14,285 (14.12)
Dyspnea	6920 (6.84)
Minor Complaint	6536 (6.46)
Dizziness	4920 (4.86)
Headache	3643 (3.60)
Laceration	2366 (2.34)
Skin Rash	2323 (2.30)
Head Trauma	2259 (2.23)
Pain (Lower Extremity)	2011 (1.99)
Chest Pain (Suspected Cardiogenic Pain)	1859 (1.84)
Injury (Upper Extremity)	1820 (1.80)
Injury (Lower Extremity)	1734 (1.71)
Pain (Upper Extremity)	1639 (1.62)
Limb Weakness	1488 (1.47)
Inter-Hospital Transfer	1434 (1.42)
Altered Mentality	1393 (1.38)
Back Pain	1317 (1.3)
Coughing and Stuffy Nose	1197 (1.18)
Palpitation and Irregular Heart Rate	1196 (1.18)
Hematochezia/Melena	1169 (1.16)
Seizure	1154 (1.14)
Nausea/Vomiting	1106 (1.09)
General Weakness	1037 (1.02)
Injury (Facial)	994 (0.98)
Chest Pain (Noncardiogenic)	860 (0.85)
Hematuria	854 (0.84)
Other	18,592 (18.37)

Override Patterns

Of the total 102,887 alerts, 65,616 (63.77%) alerts were overridden. We then analyzed the effects of physician-related factors, patient-related factors, and alert-related factors that could affect the physicians' alert override behavior (Table 4).

Table 4. The risk of alert overrides according to various factors.

Factor	Frequency of alert (n)	Alert override rate, n (%)	Univariate logistic regression odds ratio (95% CI)	Multivariate logistic regression odds ratio (95% CI)
Physician-related factors				
Physicians' Designation				
Resident	93,022	59,678 (64.15)	1 [Reference]	1 [Reference]
Fellow	8174	4993 (61.08)	0.88 (0.84-0.92)	0.9 (0.86-0.94)
Faculty	1691	945 (55.88)	0.71 (0.64-0.78)	0.73 (0.66-0.81)
Physicians' Specialty				
Emergency Department	50,812	32,542 (64.04)	1 [Reference]	1 [Reference]
Internal Medicine	17,476	10,623 (60.79)	0.87 (0.84-0.9)	1.03 (0.99-1.07)
Surgical Department	8737	5501 (62.96)	0.95 (0.91-1)	0.90 (0.85-0.94)
Other Department	25,862	16,950 (65.54)	1.07 (1.03-1.1)	1.10 (1.06-1.14)
Patient-related factors				
Patients' severity score				
1 (Highest Severity)	1597	912 (57.11)	0.80 (0.73-0.89)	0.82 (0.74-0.91)
2	8985	5421 (60.33)	0.92 (0.88-0.96)	0.89 (0.85-0.94)
3	45,759	28,536 (62.36)	1 [Reference]	1 [Reference]
4	41,171	27,059 (65.72)	1.16 (1.13-1.19)	1.07 (1.04-1.10)
5 (Lowest Severity)	5375	3688 (68.61)	1.32 (1.24-1.40)	1.23 (1.15-1.32)
Patients' chief complaints				
Fever	26,334	16,769 (63.68)	1 [Reference]	1 [Reference]
Abdominal Pain	11,300	6525 (57.74)	0.78 (0.75-0.82)	0.95 (0.91-1.00)
Altered Mentality	3220	2080 (64.6)	1.04 (0.96-1.12)	1.49 (1.37-1.62)
Back Pain	1684	1079 (64.07)	1.02 (0.92-1.13)	1.04 (0.94-1.16)
Chest Pain (Non-Cardiogenic)	680	388 (57.06)	0.76 (0.65-0.88)	0.84 (0.72-0.98)
Chest Pain (Suspected Cardiogenic Pain)	1946	1216 (62.49)	0.95 (0.86-1.05)	1.22 (1.09-1.37)
Coughing and Stuffy Nose	1195	769 (64.35)	1.03 (0.91-1.16)	1.04 (0.92-1.18)
Dizziness	3128	2098 (67.07)	1.16 (1.07-1.26)	1.65 (1.52-1.79)
Dyspnea	8432	4894 (58.04)	0.79 (0.75-0.83)	0.93 (0.88-0.98)
General Weakness	1419	909 (64.06)	1.02 (0.91-1.14)	1.28 (1.15-1.44)
Head Trauma	1327	985 (74.23)	1.64 (1.45-1.86)	1.64 (1.45-1.87)
Headache	4165	2723 (65.38)	1.08 (1.01-1.15)	1.08 (1.01-1.16)
Hematochezia/Melena	3236	2209 (68.26)	1.23 (1.13-1.33)	2 (1.84-2.18)
Hematuria	561	379 (67.56)	1.19 (1.00-1.42)	1.25 (1.05-1.50)
Injury (Facial)	556	446 (80.22)	2.31 (1.88-2.87)	2.18 (1.76-2.70)
Injury (Lower Extremity)	1335	905 (67.79)	1.2 (1.07-1.35)	1.29 (1.14-1.46)
Injury (Upper Extremity)	1103	817 (74.07)	1.63 (1.42-1.87)	1.56 (1.36-1.80)
Inter-hospital Transfer	1354	855 (63.15)	0.98 (0.87-1.10)	1.10 (0.97-1.24)
Laceration	1834	1585 (86.42)	3.63 (3.18-4.17)	3.43 (2.98-3.95)
Limb Weakness	1173	689 (58.74)	0.81 (0.72-0.91)	1.01 (0.89-1.14)
Minor Complaint	5082	3408 (67.06)	1.16 (1.09-1.24)	1.18 (1.10-1.27)
Nausea/Vomiting	921	455 (49.4)	0.56 (0.49-0.64)	0.7 (0.61-0.80)
Pain (Lower Extremity)	1800	1111 (61.72)	0.92 (0.83-1.02)	0.97 (0.88-1.08)

Factor	Frequency of alert (n)	Alert override rate, n (%)	Univariate logistic regression odds ratio (95% CI)	Multivariate logistic regression odds ratio (95% CI)
Pain (Upper Extremity)	1043	731 (70.09)	1.34 (1.17-1.53)	1.3 (1.13-1.49)
Palpitation and Irregular Heart Rate	693	404 (58.3)	0.8 (0.68-0.93)	0.98 (0.84-1.15)
Seizure	1993	1179 (59.16)	0.83 (0.75-0.91)	0.93 (0.84-1.02)
Skin Rash	1776	1141 (64.25)	1.02 (0.93-1.13)	1.18 (1.06-1.32)
Others	13,597	8867 (65.21)	1.07 (1.02-1.12)	1.29 (1.23-1.35)
Alert-related factors				
Type of alert				
Duplication	1414	911 (64.43)	1 [Reference]	1 [Reference]
Age	17,949	11,035 (61.48)	0.88 (0.79-0.99)	0.8 (0.71-0.90)
Allergy	1583	822 (51.93)	0.6 (0.51-0.69)	0.54 (0.46-0.62)
Disease	4041	2479 (61.35)	0.88 (0.77-0.99)	0.93 (0.82-1.06)
Dose	67,212	43,873 (65.28)	1.04 (0.93-1.16)	0.99 (0.88-1.11)
Drug-Drug Interaction	1399	926 (66.19)	1.08 (0.93-1.26)	1.07 (0.91-1.25)
Gender	381	183 (48.03)	0.51 (0.41-0.64)	0.43 (0.33-0.56)
Lactation	18	13 (72.22)	1.44 (0.54-4.50)	1.33 (0.50-4.18)
Pregnancy	8651	5199 (60.10)	0.83 (0.74-0.93)	0.72 (0.64-0.81)
Route	239	175 (73.22)	1.51 (1.12-2.06)	1.19 (0.88-1.64)
Medication categories (based on generic product ID)				
Neuromuscular Drugs	1511	961 (63.60)	1 [Reference]	1 [Reference]
Analgesics and Anesthetics	36,685	24,808 (67.62)	1.20 (1.07-1.33)	1.23 (1.10-1.38)
Anti-Infective Agents	18,308	12,419 (67.83)	1.21 (1.08-1.35)	1.08 (0.96-1.21)
Antineoplastic	326	198 (60.74)	0.89 (0.69-1.13)	0.94 (0.73-1.22)
Biologicals	192	138 (71.88)	1.46 (1.06-2.05)	1.09 (0.78-1.53)
Cardiovascular Agents	5866	3637 (62)	0.93 (0.83-1.05)	0.95 (0.84-1.07)
Central Nervous System Drugs	5434	3066 (56.42)	0.74 (0.66-0.83)	0.67 (0.59-0.76)
Endocrine and Metabolic Drugs	3374	2009 (59.54)	0.84 (0.74-0.95)	0.84 (0.74-0.96)
Gastrointestinal Agents	15,589	8619 (55.29)	0.71 (0.63-0.79)	0.6 (0.53-0.67)
Genitourinary Agents	343	211 (61.52)	0.91 (0.72-1.17)	1.42 (1.07-1.90)
Hematological Agents	4938	3210 (65.01)	1.06 (0.94-1.20)	1.04 (0.91-1.18)
Miscellaneous Products	1168	711 (60.87)	0.89 (0.76-1.04)	0.92 (0.78-1.09)
Nutritional Product	1843	1214 (65.87)	1.1 (0.96-1.27)	1.09 (0.94-1.26)
Respiratory Agents	6778	3994 (58.93)	0.82 (0.73-0.92)	0.85 (0.75-0.96)
Topical Products	532	421 (79.14)	2.17 (1.72-2.75)	1.62 (1.27-2.07)

In terms of physician-related factors, the resident group had a higher override rate than both the fellow and faculty groups. The top three physicians' specialty departments that generated the most alerts in the ED were the emergency medicine, pediatrics, and general internal medicine departments. Physicians working in the ED were found to override over 64% of the total alerts received. In terms of patient factors, the alert override rate tended to decrease with the increase in severity. Additionally, the alert override rate also showed wide variation according to the patients' chief complaints. The override rate tended to be higher in patients with trauma such as laceration and injuries than in patients with other chief complaints.

Two-thirds of the total alerts were dose alerts, and 65.3% of these were overridden. ED physicians overrode approximately half of the gender- and allergy-type alerts. Regarding the medication group, alerts for gastrointestinal agents, central nervous system drugs, respiratory agents, and endocrine and metabolic drugs showed the lowest override rates. Among the variables used in multivariate logistic regression models, the following variables emerged as statistically significant risk

factors: miscellaneous department group in physician's specialty; patients with lower and lowest severity; the presence of dizziness, head trauma, headache, hematochezia, melena, hematuria, facial injuries, lower extremity injuries, upper extremity injuries, laceration, minor complaints, and upper extremity pain; drug-drug interaction, lactation, and the route; and medication categories of analgesics and anesthetics, anti-infective agents, and topical products (Table 4).

Discussion

Principal Results

In this study, we examined the alert override patterns in an ED with a CDSS that was designed in a minimally interruptive way. Approximately two-thirds of the alerts were overridden, which is a lower rate than reported in many previous related studies. We assessed many covariates, including both physician-related and patient-related factors, as well as alert-related factors. These results could be used for optimizing and maximizing the effectiveness of the CDSS.

Physicians' Specialty, Patients' Severity, and Alert Override Rate

Many studies examining alert override rates have not examined the physicians' designations as an input variable or did not find it to have a significant effect [14,24]. Despite lack of evidence, it is generally considered that experienced physicians do not need alerts because "they already know" [25], and even if they receive them, they may be more likely to override them. In contrast to this expectation, we found that the alert override rate was the highest among residents, followed by fellows and senior faculty members (Table 4). This finding establishes that even an experienced physician still requires assistances from a CDSS.

We also demonstrated that physicians override more alerts in patients with complaints of lower severity. The override rate for patients with the highest severity level was 57.1% and that for patients with the lowest severity level was 63.5%. The override rate decreased significantly as the patients' severity increased (Table 4). To date, relatively little attention has been paid to patients' severity as a covariate affecting physicians' alert overriding behavior. Further investigation is needed to confirm this finding.

Comparison With Prior Work

Low Override Rate

One of the major findings in this study is the relatively low override rate. The low override rate was observed consistently across the three factors (physician, patient, and alert), implying that a systematic influence exists aside from those discussed above. One potential explanation could be related to the phenomenon that has been termed the "cloud of context." Coiera et al [26] proposed that "variations in the workflow, patient population and morbidity, resources, pre-existing infrastructure, and the education and experience of both clinical staff and patients" function as a "cloud of context" that affect how physicians respond to a CDSS. Given that most of the existing CDSS studies were conducted in inpatient and outpatient settings in the United States, these contextual factors may have influenced the lower override rate observed in this study. Previous studies have reported that override rates increased from 72.8% to 93% [7-10,24]. In comparison, a Korean study reported a rate of 71.7% [14], which is relatively lower than that reported for the United States. However, there are no other ED-based studies that we are aware of for direct comparison; thus, the conclusion warrants further investigation.

Another possible explanation is the system design. It is well known that utilizing a human factors design reduces the error rate and that an interactive design can reduce alert fatigue [15,27]. We leveraged several strategies to ensure that the system is integrated into the clinical workflow by designing a noninterruptive system. For example, the concept of timely alerts was implemented more seamlessly by designing a system that can generate an alert whenever a new order component is entered. Although our research scope did not include direct measurements and analysis of usability, we believe that integrating the system into the clinical workflow contributes to the override rate for ED physicians. We also believe that the underlying knowledge used was relatively robust.

Importance of the ED Environment in Alert Override–Related Research

In this study, we assessed the alert overriding patterns of ED physicians in routine care in the ED. An ED is an optimal environment to examine physicians' alert overriding behaviors from a broad perspective because patients have a wider range of severity than those in other departments, and many receive interdisciplinary care in this environment. This viewpoint is important because the CDSS does not influence physicians concerning a single type of alert but rather acts as a bundle of user interfaces.

To the best of our knowledge, this is the first study to analyze physicians' CDSS usage patterns in an ED comprehensively. In a recent review, less than one-tenth of the studies were found to target physicians' behavior in the ED [15]. Some research studies only targeted specific alert types such as drug-drug interaction or opioid alerts, or only a specific group such as pediatric patients [10,17,18]. In this study, the analysis was performed on the whole system rather than focusing on particular types of alerts, physicians, or patients. The effect of a CDSS may be critical in this complex environment, and the results of this study may play an essential role in establishing CDSSs in EDs or in cases in which a multidisciplinary approach is needed.

Limitations

First, this study was performed in a single ED with a homegrown EHR. The ED part of an academic referral center receives patients with conditions of high severity, and the majority of physicians are trainees of its residency program. Furthermore, the EHR, DARWIN, has only been implemented in a few hospitals in Korea to date, which should be considered while generalizing our results.

Second, the alert override was the only outcome measured. The outcome of the alert override or the reason for override was not recorded in the database. Thus, it was not possible to determine the clinical appropriateness of the alerts or overrides, or the

consequences for patients. Therefore, there is a need for a follow-up interview study or an institution-level investigation of adverse drug events.

Third, the effect of the minimally interrupted CDSS was not demonstrated in this ED. For clarity, a comparative study is required. Such a comparison would require a control group with more interruptive alerts on similar targets. The nature of the CDSS makes it difficult to perform a randomized controlled trial.

Finally, we did not scale all potential factors related to alert overrides in previous studies, such as alert fatigue, alert severity, and workload. In particular, to estimate the size of the effects of alert fatigue, there is a need for further observational studies using devices such as eye trackers that can quantitatively measure whether the physician paid attention to the CPOE alert.

Conclusions

In this retrospective study, we described alert override patterns with a medication CDSS in an academic ED. We found a relatively low rate of overrides and also assessed the influence of multiple contributing factors on these rates. This study could aid CDSS implementers by providing knowledge regarding physicians' alert overriding behaviors as well as empirical evidence that contradicts conventional notions.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C0275). We would like to thank Editage (www.editage.co.kr) for English language editing.

Authors' Contributions

JY collected, coded, and analyzed the data, and wrote the manuscript. JL reviewed sample alerts and inspected the manuscript. PR developed the CDSS and reviewed the manuscript. DC, MK, and JC implemented the CDSS and reviewed the manuscript. DB analyzed and inspected the manuscript. WC developed and implemented the CDSS, designed the study, oversaw data analysis, and cowrote the manuscript.

Conflicts of Interest

None declared.

References

1. Bernstein S, Aronsky D, Duseja R, Epstein S, Handel D, Hwang U, Society for Academic Emergency Medicine, Emergency Department Crowding Task Force. The effect of emergency department crowding on clinically oriented outcomes. Acad Emerg Med 2009 Jan;16(1):1-10. [doi: 10.1111/j.1553-2712.2008.00295.x] [Medline: 19007346]
2. Hafner JW, Belknap SM, Squillante MD, Bucheit KA. Adverse drug events in emergency department patients. Ann Emerg Med 2002 Mar;39(3):258-267. [doi: 10.1067/mem.2002.121401] [Medline: 11867978]
3. Hohl CM, Badke K, Zhao A, Wickham ME, Woo SA, Sivilotti MLA, et al. Prospective Validation of Clinical Criteria to Identify Emergency Department Patients at High Risk for Adverse Drug Events. Acad Emerg Med 2018 Sep;25(9):1015-1026. [doi: 10.1111/acem.13407] [Medline: 29517818]
4. Karpov A, Parcero C, Mok CP, Panditha C, Yu E, Dempster L, et al. Performance of trigger tools in identifying adverse drug events in emergency department patients: a validation study. Br J Clin Pharmacol 2016 Oct;82(4):1048-1057. [doi: 10.1111/bcp.13032] [Medline: 27279597]
5. Gerrity MS, DeVellis RF, Light DW. Uncertainty and Professional Work: Perceptions of Physicians in Clinical Practice. Am J Sociol 1992 Jan;97(4):1022-1051. [doi: 10.1086/229860]
6. Shojania KG, Duncan BW, McDonald KM, Wachter RM, Markowitz AJ. Making health care safer: a critical analysis of patient safety practices. Evid Rep Technol Assess (Summ) 2001(43):i-x, 1-668. [Medline: 11510252]
7. Wong A, Rehr C, Seger DL, Amato MG, Beeler PE, Slight SP, et al. Evaluation of Harm Associated with High Dose-Range Clinical Decision Support Overrides in the Intensive Care Unit. Drug Saf 2019 Apr;42(4):573-579. [doi: 10.1007/s40264-018-0756-x] [Medline: 30506472]
8. Slight SP, Beeler PE, Seger DL, Amato MG, Her QL, Swerdloff M, et al. A cross-sectional observational study of high override rates of drug allergy alerts in inpatient and outpatient settings, and opportunities for improvement. BMJ Qual Saf 2017 Mar;26(3):217-225 [FREE Full text] [doi: 10.1136/bmjqs-2015-004851] [Medline: 26993641]
9. Nanji KC, Seger DL, Slight SP, Amato MG, Beeler PE, Her QL, et al. Medication-related clinical decision support alert overrides in inpatients. J Am Med Inform Assoc 2018 May 01;25(5):476-481. [doi: 10.1093/jamia/ocx115] [Medline: 29092059]
10. Isaac T, Weissman JS, Davis RB, Massagli M, Cyrulik A, Sands DZ, et al. Overrides of medication alerts in ambulatory care. Arch Intern Med 2009 Feb 09;169(3):305-311. [doi: 10.1001/archinternmed.2008.551] [Medline: 19204222]

11. Shekelle P, Wachter R, Pronovost P, Schoelles K, McDonald K, Dy S, et al. Making health care safer II: an updated critical analysis of the evidence for patient safety practices. Evid Rep Technol Assess (Full Rep) 2013 Mar(211):1-945. [Medline: 24423049]
12. Prgomet M, Li L, Niazkhani Z, Georgiou A, Westbrook JI. Impact of commercial computerized provider order entry (CPOE) and clinical decision support systems (CDSSs) on medication errors, length of stay, and mortality in intensive care units: a systematic review and meta-analysis. J Am Med Inform Assoc 2017 Mar 01;24(2):413-422. [doi: 10.1093/jamia/ocw145] [Medline: 28395016]
13. Ranji SR, Rennke S, Wachter RM. Computerised provider order entry combined with clinical decision support systems to improve medication safety: a narrative review. BMJ Qual Saf 2014 Sep;23(9):773-780. [doi: 10.1136/bmjqs-2013-002165] [Medline: 24728888]
14. Cho I, Lee Y, Lee J, Bates DW. Wide variation and patterns of physicians' responses to drug-drug interaction alerts. Int J Qual Health Care 2019 Mar 01;31(2):89-95. [doi: 10.1093/intqhc/mzy102] [Medline: 29741633]
15. Hussain MI, Reynolds TL, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. J Am Med Inform Assoc 2019 Oct 01;26(10):1141-1149 [FREE Full text] [doi: 10.1093/jamia/ocz095] [Medline: 31206159]
16. Tolley C, Forde N, Coffey K, Sittig D, Ash J, Husband A, et al. Factors contributing to medication errors made when using computerized order entry in pediatrics: a systematic review. J Am Med Inform Assoc 2018 May 01;25(5):575-584. [doi: 10.1093/jamia/ocx124] [Medline: 29088436]
17. Genco EK, Forster JE, Flaten H, Goss F, Heard KJ, Hoppe J, et al. Clinically Inconsequential Alerts: The Characteristics of Opioid Drug Alerts and Their Utility in Preventing Adverse Drug Events in the Emergency Department. Ann Emerg Med 2016 Feb;67(2):240-248.e3 [FREE Full text] [doi: 10.1016/j.annemergmed.2015.09.020] [Medline: 26553282]
18. Sethuraman U, Kannikeswaran N, Murray KP, Zidan MA, Chamberlain JM. Prescription errors before and after introduction of electronic medication alert system in a pediatric emergency department. Acad Emerg Med 2015 Jun;22(6):714-719. [doi: 10.1111/acem.12678] [Medline: 25998704]
19. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012 Oct 08;172(18):1377-1385. [doi: 10.1001/archinternmed.2012.3199] [Medline: 22911330]
20. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
21. Arora M, Asha S, Chinnappa J, Diwan AD. Review article: burnout in emergency medicine physicians. Emerg Med Australas 2013 Dec;25(6):491-495. [doi: 10.1111/1742-6723.12135] [Medline: 24118838]
22. Jung KY, Kim T, Jung J, Lee J, Choi JS, Mira K, et al. The Effectiveness of Near-Field Communication Integrated with a Mobile Electronic Medical Record System: Emergency Department Simulation Study. JMIR Mhealth Uhealth 2018 Sep 21;6(9):e11187 [FREE Full text] [doi: 10.2196/11187] [Medline: 30249577]
23. Choi H, Ok JS, An SY. Evaluation of Validity of the Korean Triage and Acuity Scale. J Korean Acad Nurs 2019 Feb;49(1):26-35. [doi: 10.4040/jkan.2019.49.1.26] [Medline: 30837440]
24. Cho I, Slight SP, Nanji KC, Seger DL, Maniam N, Fiskio JM, et al. The effect of provider characteristics on the responses to medication-related decision support alerts. Int J Med Inform 2015 Sep;84(9):630-639. [doi: 10.1016/j.ijmedinf.2015.04.006] [Medline: 26004341]
25. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13(2):138-147 [FREE Full text] [doi: 10.1197/jamia.M1809] [Medline: 16357358]
26. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? J Am Med Inform Assoc 2018 Aug 01;25(8):963-968 [FREE Full text] [doi: 10.1093/jamia/ocy028] [Medline: 29669066]
27. Brown CL, Mulcaster HL, Triffitt KL, Sittig DF, Ash JS, Reygate K, et al. A systematic review of the types and causes of prescribing errors generated from using computerized provider order entry systems in primary and secondary care. J Am Med Inform Assoc 2017 Mar 01;24(2):432-440. [doi: 10.1093/jamia/ocw119] [Medline: 27582471]

Abbreviations

CDSS: computerized decision support system
CPOE: computerized provider order entry
DARWIN: Data Analytics and Research Window for Integrated Knowledge
ED: emergency department
EHR: electronic health record
KTAS: Korean Triage and Acuity Scale

Edited by G Eysenbach; submitted 09.08.20; peer-reviewed by A Azzam, P Beeler, S Sarbadhikari; comments to author 31.08.20; revised version received 12.09.20; accepted 21.10.20; published 04.11.20.

Please cite as:
Yoo J, Lee J, Rhee PL, Chang DK, Kang M, Choi JS, Bates DW, Cha WC
Alert Override Patterns With a Medication Clinical Decision Support System in an Academic Emergency Department: Retrospective Descriptive Study
JMIR Med Inform 2020;8(11):e23351
URL: https://medinform.jmir.org/2020/11/e23351
doi:10.2196/23351
PMID:33146626

Viewpoint

Amplifying Domain Expertise in Clinical Data Pipelines

Protiva Rahman[1], PhD; Arnab Nandi[1], PhD; Courtney Hebert[1], MSc, MD

The Ohio State University, Columbus, OH, United States

Corresponding Author:
Protiva Rahman, PhD
The Ohio State University
1800 Cannon Drive
Columbus, OH, 43210
United States
Phone: 1 3056135087
Email: rahman.92@osu.edu

Abstract

Digitization of health records has allowed the health care domain to adopt data-driven algorithms for decision support. There are multiple people involved in this process: a data engineer who processes and restructures the data, a data scientist who develops statistical models, and a domain expert who informs the design of the data pipeline and consumes its results for decision support. Although there are multiple data interaction tools for data scientists, few exist to allow domain experts to interact with data meaningfully. Designing systems for domain experts requires careful thought because they have different needs and characteristics from other end users. There should be an increased emphasis on the system to optimize the experts' interaction by directing them to high-impact data tasks and reducing the total task completion time. We refer to this optimization as amplifying domain expertise. Although there is active research in making machine learning models more explainable and usable, it focuses on the final outputs of the model. However, in the clinical domain, expert involvement is needed at every pipeline step: curation, cleaning, and analysis. To this end, we review literature from the database, human-computer information, and visualization communities to demonstrate the challenges and solutions at each of the data pipeline stages. Next, we present a taxonomy of expertise amplification, which can be applied when building systems for domain experts. This includes summarization, guidance, interaction, and acceleration. Finally, we demonstrate the use of our taxonomy with a case study.

(JMIR Med Inform 2020;8(11):e19612) doi:10.2196/19612

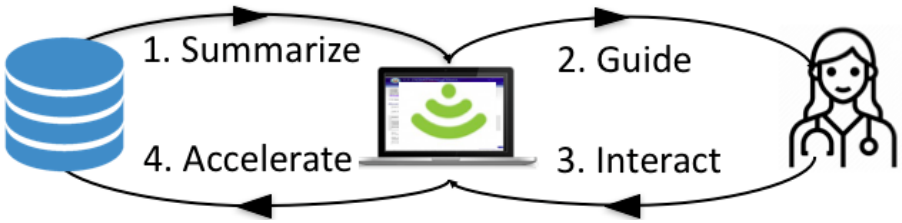
KEYWORDS

review; data analysis; data science; clinical informatics

Introduction

Recent advancements in data availability (eg, digitization of health records) and deep neural networks [1] have led to the resurgence of artificial intelligence. This has served as a catalyst for data-driven decision making in many domains. However, for high-stakes applications, such as financial and health care domains, it is rare for domain experts to execute decisions solely based on artificial intelligence algorithms [2]. Domain experts in this context are individuals who are not necessarily trained in computational fields but inform the design and are end users of data-driven algorithms (eg, health care providers, hospital administrators). Note that domain experts can have different levels of expertise in their specific domain (eg, interns, residents, attendings), and we do not differentiate between these levels in this work. Although the role of experts has been studied in clinical decision support (CDS), we find a gap in their involvement in the data analysis pipeline, which we focus on in this work.

Figure 1. Domain expertise amplification.



Domain expert involvement remains necessary in the health care domain, but this involvement brings significant challenges and implications for data-driven applications. Domain experts are expensive resources with limited time for these efforts, and excessive reliance on domain expertise could potentially lead to systems that are overly customized and not reproducible or scalable. Owing to these challenges, designing systems for them requires careful consideration. To address these challenges, we present a framework for amplified intelligence that identifies the points in the process where expertise can be effectively leveraged. *Amplification of expertise* then refers to the process of automating redundant or inferable tasks, so that domain experts can focus their efforts on tasks that require domain knowledge. This is a synergy between the domain expert and the system, which involves summarization of data and decisions, guidance toward insights, interaction by the domain expert, and acceleration of input (Figure 1).

Prior Work

There is active research on interactive and human-in-the-loop systems in many computer science disciplines. The database and visualization communities have produced numerous tools [3-8] to aid data scientists with data wrangling and analysis. At the decision-making stage, the machine learning community has looked at making black box models explainable [2,9-12], while the human-computer interaction (HCI) community has been studying how differences in explainability affect decision making [13,14]. Finally, the crowdsourcing community has concentrated on human-powered computation by optimizing tasks (eg, simplifying tasks [15], minimizing the number of questions [16,17], optimizing workflows [18-20]). However, we focus on data-powered experts by amplifying expertise. Although we draw from prior work, systems designed for health care domain experts require special consideration because they have characteristics that distinguish them from data scientists and crowdworkers.

Special Considerations in the Health Care Domain

First, domain expert input is usually needed for data tasks that require experiential knowledge and judgment (such as medical diagnoses and forensic analysis [21]). The critical and subjective nature of these decisions necessitates transparency, both from the algorithm and domain experts. Hence, the system needs to summarize the impact of algorithmic or experts' manipulation of the data [22]. Second, due to their specialized training, domain experts' time is expensive and limited [23,24]. This constraint makes it imperative that we build tools that provide insights while reducing physical and cognitive effort [25]. Third, as domain experts are trained in noncomputational fields, systems designed for them should provide high-level interaction capabilities. This is referred to as *editable shared representations* between computers and humans [26]. Examples include natural language interfaces and form-based input [27]. Finally, domain experts are highly trained individuals, which allows systems to accelerate their input by using domain-specific assumptions and ontologies [28,29]. Keeping these factors in mind, expertise amplification involves summarization, guidance, interaction, and acceleration (Figure 1). We will explore each of these in detail in the following sections.

The Data Pipeline

There are opportunities to amplify expertise at all stages of the pipeline. The data pipeline refers to the different stages that the data need to go through before they can provide decision support. It can roughly be broken into 3 stages: curation, cleaning, and analysis. Tools at the end of the pipeline have only looked at explaining models but not at amplification. In contrast, tools at earlier pipeline stages have been designed mainly for data scientists and not for experts. However, domain experts are involved at every stage of the pipeline [27-31], especially in clinical research settings where data sets contain specialized information. Thus, there is a need to amplify domain expertise throughout the pipeline. In this work, we provide examples from the informatics literature to highlight the need for expert involvement at each pipeline step. We then review literature from the database, HCI, and visualization communities about challenges and current approaches at different stages. On the basis of our review, we present a novel taxonomy for amplifying domain expertise and demonstrate its use with a case study in empiric antibiotic treatment. Our review can serve as a guide to new clinical research projects, and our taxonomy can be applied when designing systems for experts, especially for low-budget projects when there are limited resources and availability of domain experts.

Challenges in the Data Pipeline

This section is organized to reflect the clinical data pipeline, which often involves the following steps: data are curated from the electronic health record (EHR) data warehouse and annotated with external data sources, cleaned and validated, and analyzed. Multiple people are involved in various stages of the pipeline. The prevalent notion of the workflow is that a data engineer restructures, cleans, and sets up the infrastructure for data analysis, and a data scientist then analyzes and models the data, which a software engineer implements into a decision support system. A domain expert then consumes the end product to make decisions. However, in clinical settings, domain expert involvement is required at every step of the pipeline. Allowing domain experts to directly and efficiently interact with data removes the need for them to rely on a data engineer or data scientist who can then focus on infrastructure and model construction. Moreover, since domain experts are the stakeholders in the output of data pipelines, in our experience, they tend to be engaged users who want to interact with data and leverage their expertise. In this section, we motivate domain expert involvement with examples from the past five years of research presented at the American Medical Informatics Association's annual symposiums. We then review the computer science literature to identify current tools and opportunities for expertise amplification at the 3 stages of the data pipeline: data curation, data cleaning, and data analysis, as each of these corresponds to a research area of its own.

Data Curation

Curating data sets for analysis can be a laborious process that can involve combining multiple data sources and identifying relevant attributes. Data integration and data discovery address these problems.

Data Integration

Medical data pipelines often involve data that were collected for purposes other than answering the research question at hand. This usually implies that information is not captured in a manner fit for analysis [32,33], with issues such as missing metadata information [34]. Moreover, in some situations such as rare disease studies, the cohort size is too small for analysis [35], while in other cases, external features such as air quality or drug components [36-39] might be needed. One possible solution to these data quality issues is to curate data from multiple institutions and external sources. However, the different data representations [35,40] pose challenges in entity matching, metadata inference, and data integrity [41,42]. Data integration aims to automatically resolve schema matching and entity matching problems during data curation. For biomedical data sets, integration can involve standardization by mapping to ontologies with controlled vocabularies [43-45]. Although current approaches use deep learning for integration [46-50], generating a training corpus and validating results require domain expert input. For example, Cui et al [35] require domain experts to validate data curation efforts for studying sudden death in epilepsy. In another example, building an automatic concept annotator for standardizing biomedical literature [50] required experts to manually annotate different concepts [51-54]. Furthermore, a domain expert will be able to catch inconsistencies or errors made by an automated integration tool much faster than a data engineer who is unfamiliar with the domain. Thus, there is a need to build interactive data integration tools for domain experts.

Data Discovery

Data discovery refers to the process of finding relevant attributes or cohorts for analysis. This is especially true for multidisciplinary teams where the domain expert knows the disease definition but is not familiar with the database schema. At the same time, the data engineer can explore the schema but might not recognize that a field is relevant. Integrating data from multiple sources only exacerbates this problem. In the informatics community, DIVA [55] aids in cohort discovery by ingesting expert-defined constraints, while visual analytic systems [56,57] such as CAVA provide an interactive interface. In the database community, Nargesian et al [58,59] have looked at finding unionable (more data points) and joinable (more attributes) data for a given data set. These algorithms are useful when trying to augment data sets with publicly available data sets such as MIMIC [60] or even for exploring a complex schema such as the Unified Medical Language System (UMLS) [61]. In addition to using properties of the data to find possible attribute matches, domain rules can be useful for identifying relevant data subsets. This requires an interactive interface where domain experts can look at subsets of interest and iteratively join and filter the data [62] to find the required cohort. Recently, query logs have been used to design precision interfaces [63,64] that customize the interface for the user's task.

Data Cleaning

After curating relevant data sets, data still need to go through multiple preprocessing steps before they are analysis-ready. These include identifying and fixing incorrect data, data augmentation, and data transformation [65], all of which benefit from domain expert involvement.

Error Fixes

EHR data are known to be messy and have errors and missing values [66-68]. A typical data cleaning method is the use of rule-based systems that identify dirty data by detecting violations of user-specified rules or known functional dependencies [69-78]. These systems do not optimize the expert's rule specification process. Crowdsourcing systems have also been used to correct values [18,79], although they are not always an option due to data complexity or confidentiality. Another approach to identify and clean data is to augment the data with external knowledge bases [80-82]. More recently, there have been many approaches [83-85] that use deep learning for automated data cleaning. Of note is Holoclean [84], which uses a statistical model to combine various data repair signals such as violation of integrity constraints, functional dependencies, and knowledge bases. Although this achieves higher performance than using each method in isolation, there is scope for identifying which of the signals are performing the poorest or what additional information would help improve the system's performance. Identifying this information, incorporating domain knowledge, and presenting it succinctly to a domain expert remain open problems.

Data Augmentation

Although data entry errors [86] and missing information can be imputed by semiautomated methods, a more difficult problem is that of creating a gold standard for training data, which is referred to as data augmentation. Many health care applications require annotating training data, for example, clinical text annotation [87-89], CDS [90-92], identifying new terms for ontologies [93], index terms for articles [94], and disease-specific annotations [51,95,96]. However, very few applications focus on optimizing the domain expert's data augmentation effort, which is eventually crucial to model performance. A notable approach to this is the Snorkel system [97], which automates data augmentation by learning the labeling function, thus accelerating the domain expert's input. However, there are opportunities to make the initial labeling process more interactive, as domain experts are required to write code in Snorkel. Furthermore, the system does not provide feedback on how labels affect the data set or final model, which is crucial for building trust in medical pipelines. Examples of interactive solutions include Icarus [28] for augmenting microbiology data and Halpern et al's system [98] for annotating clinical anchors. Both systems use an ontology to interactively amplify domain expertise.

Data Transformation

Other than fixing incorrect values and augmenting data sets, often, data need to be restructured (eg, splitting values in a column, reformatting dates). Data wrangling has emerged as a separate field in the past decade because of data diversity. Potter's Wheel [99] is one of the first interactive data transformation systems. It allows the user to specify data transforms that are encoded as constraints and used to detect errors. Building on this idea, systems such as Polaris [100] and

Trifacta [4,101] infer syntactic rules from user edits. Similarly, programming-by-example systems [102,103] learn transformations from a set of input-output pairs. These techniques have informed the autofill function of Microsoft Excel. As many domain experts employ Excel for data transformations and analysis [104], spreadsheet interfaces should consider incorporating domain knowledge.

Data Analysis

We now move to the final step of the pipeline. This includes exploratory analysis to identify attributes of interest and explainability of models for decision making.

Data Exploration

During the exploration step, it is crucial for the domain expert to be able to directly interact with the data for effective hypothesis generation. However, domain experts often must go through a data engineer to execute the relevant query [105,106] or extract information from unstructured notes [107]. The data are then validated by the domain expert through manual chart review, since data engineers without domain knowledge may apply naive filters that hide insights or find spurious correlations. To address these challenges, the informatics community has built tools to accelerate chart review [108] and allow interactive filtering and analysis [109,110]. Finalizing an analysis data set can then take multiple iterations of requests and validations between the domain expert and data engineer. In some cases, data engineers create custom dashboards for domain experts [111-113], but the latter are then limited to brushing and linking on the provided view. Mixed-initiative interfaces such as Tableau [100] and Dive [5] recommend visualizations based on statistical properties of the data but do not use domain-specific ontologies that can enrich the domain experts' interaction and accelerate their workflow.

Visualizations, when used appropriately, can provide effective summaries and reveal patterns not immediately evident by statistical overviews [114]. Summaries reduce the cognitive load on domain experts during multidimensional data exploration, allowing them to drill down to specific instances as needed [115]. Although many visualization recommendation systems exist for analyzing numerical data [7,116-118], visualizations in health care often include categorical and text data [119-122]. As such, node-link diagrams are a common data representation and have been used for tracking family history [123], decision making [22,124], and identifying hidden variables [125]. Visual interfaces thus amplify expertise by summarizing data. However, they can be more powerful if they allow interaction, provide guidance by highlighting interesting regions for exploration [126], and accelerate workflows by extrapolating domain expert interactions based on properties of the data [22]. Thus, there is a need to provide domain experts with tools that allow for more sophisticated data interaction.

Explainability

Finally, we cannot discuss clinical pipelines without discussing explainability. The interpretability of rule-based systems has made them popular in a variety of clinical applications, including decision support [127,128], antibiotic recommendation [129], updating annotations [130], and auditing [131]. Interpretability is essential because domain experts want a cause-and-effect relationship, based on which actionable decisions can be made [66,68,132]. Furthermore, health care providers may not use models they do not trust, and building trust requires providing context and explanations [2].

Current approaches in health care research use weights and activation of features to characterize attribute importance [133-135]. RuleMatrix [136] provides an alternate approach where a set of rules represents the deep learning model. The expert can explore various facets of each rule, such as data affected, distribution, and errors. In another example, Cai et al [29] built a tool to help pathologists find similar images to aid in diagnoses. The tool allows domain experts to search for similar images and then interactively refine the search results. It allows refinement by region (crop an image), refinement by concept (filter by extracted concepts from image embeddings), and refinement by example (select multiple images as examples). These refinement techniques are examples of acceleration, where interactions are interpolated to the entire data set by learning general functions. Explainability is thus key to the adoption of deep learning models. Although they have mainly been applied in the analysis stage of the pipeline, they are equally important when applying automated algorithms to curation and cleaning.

Therefore, amplifying domain experts' abilities in the analysis stage requires interactive data systems using a combination of statistical algorithms and compelling visualizations. Moreover, these systems need to follow design-study principles [137]. They need to allow interaction with domain experts for a needs assessment and an empirical evaluation to ensure that correct information is portrayed effectively. Otherwise, the system can end up burdening and biasing the domain expert instead of helping [13,101].

We have highlighted the need for domain expert involvement in the pipeline and describe some of the challenges they encounter. Although we have briefly expanded on some available solutions, Table 1 provides a more comprehensive list of references. Summarizing each technique is outside the scope of this paper, but it provides a guide to interested readers for further reading.

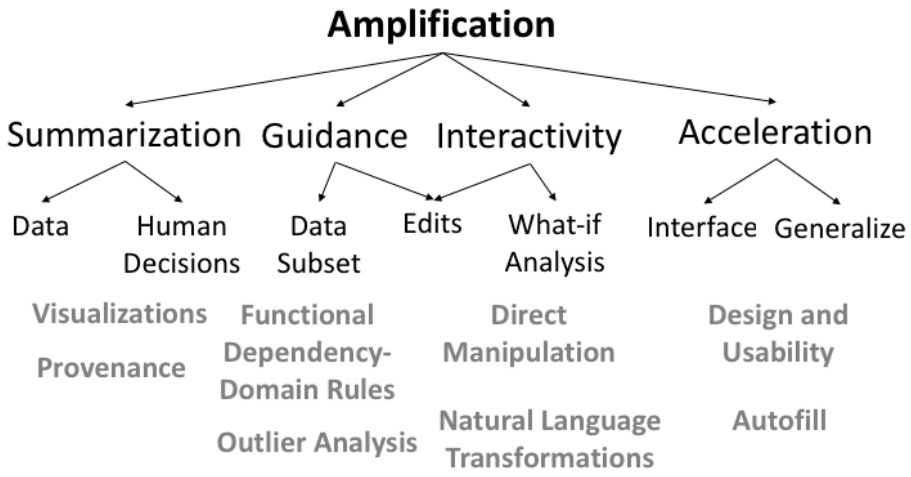
Table 1. Review of current approaches for each data pipeline stage.

Current solutions	Domain expert role
Data curation	
Data integration	
• Schema matching [138-143] • Interactive integration [144,145] • Webtables integration [146-151] • Machine learning [46-49]	Domain experts are needed to validate results of integration, and interactively correct automated methods, which can then update their algorithm
Data discovery	
• Attribute search [58,59,152,153] • Interactive querying [55,62-64] • Visual analytics [56,57]	Domain expert feedback is needed to finalize the analysis data set
Data cleaning	
Error fixes	
• Rule-based [69-77,154] • Crowdbased [18,79,155,156] • Knowledgebase [80-82] • Machine learning [83-85] • Functional dependency [15-23,25-54,58-165]	Domain expert input can be used to identify and fix errors
Augmentation	
• Machine learning [97,166,167] • Interactive [28,98,168-170]	Domain experts can augment missing data with domain-specific rules
Transformation	
• Programming by example [102,103] • Interactive rules [4,99-101] • Foreign-key detection [153,171-175]	Domain experts can restructure the data to make it semantically valid
Data analysis	
Exploration	
• Optimize performance [176-179] • Optimize insight [126,180,181] • Provenance [182,183] • Visualizations [5,7,116-118,184-188]	Domain experts interact with summaries and outliers to draw insight
Explainable	
• Systems [189,190] • Visualizations [9,12,29,136] • Empirical studies [10,11,13,14]	Domain experts inform the model design to ensure explainability

Taxonomy of Expertise Amplification

The previous section elucidated the need for domain expert involvement throughout the clinical data pipeline. In all steps, domain expert involvement can improve automated methods but must be implemented appropriately to ensure that the process remains robust and reproducible. Taking this into consideration, we propose a taxonomy that can be employed when designing systems to amplify expertise in the clinical pipeline. Domain expertise amplification by a system can broadly be categorized into 4 dimensions: summarization, guidance, interactivity, and acceleration, as shown in Figures 1 and 2. Thus, a system that wishes to amplify expertise should apply one or more of these dimensions. We demonstrate these categories with examples from computer science literature.

Figure 2. Taxonomy of expertise amplification: the first level shows the 4 dimensions that should be employed by a system for expertise amplification. The second level enumerates the subdimensions along which amplification can be done, while the fourth level in gray shows tools that can be applied.



Summarization

The time constraints of experts along with transparency requirements in the clinical domain motivate the need for effective summaries of data and human decisions. Although data summaries are important for analysis, summaries of human decisions allow for improved explainability and reproducibility.

Data

An amplification system should summarize large and complex data sets so that experts can meaningfully consume them. This is relevant for identifying inconsistencies as well as for open-ended exploration during analysis. It can be overwhelming for an expert to go through large and wide tables. Therefore, amplification systems should automatically summarize complex data [191]. Although providing data samples [28,76] and statistical summaries such as mean, variance, and standard deviation can be useful for providing a bird's eye view, they are not always enough to reveal patterns [114]. In such cases, *visual summaries* can provide additional insight, as done by the CAVA system [56]. Multidimensional data can be visually summarized by presenting each dimension as a coordinated histogram with linked brushing and filtering [176].

Human Decisions

In addition to data, amplification systems need to summarize algorithmic and human decisions as well. This is because domain expert involvement is usually required in situations where it is necessary to have high-quality data [2,21]. Hence, amplification systems also require high transparency [189,192]. To support algorithm transparency, amplification systems can show visual activation of features that led to the recommendation [9] or similar cases in the data that serve as evidence for the current recommendation [193]. Summarizing human decisions can involve expressing data transformations as natural language rules [4,28] and visual node-link diagrams [22]. Furthermore, as summarized data provide an abstract or aggregate view, there is a need for data transparency, meaning that experts should be able to trace individual data points, which contributed to the aggregate summary. This involves incorporating ideas from *provenance systems* such as Smoke [182] and Scorpion [183], which provide fast data lineage tracking. Finally, for each application, empirical studies are needed to see what and how information should be presented or summarized because too much transparency can overwhelm and negatively impact the expert [13].

Guidance

Although summaries provide a global view of the data, the goals of exploratory analysis include finding insights and data quality issues [191], which might require looking at a more detailed view. Systems can guide experts by navigating to informative subsets and by suggesting data transformations and edits.

Data Subset

Amplification systems should guide the expert's navigation to meaningful subsets. For example, SeeDB [116] automatically finds interesting visualizations. Given a query, it defines *interestingness* as the deviation of the query's result set from a baseline data set. In a similar vein**,** TPFlow [194] uses tensor decomposition to guide users to interesting regions in spatiotemporal exploration. For data cleaning, error detection algorithms such as Uguide [78] and DataProf [76] use *functional dependencies* and Armstrong samples, respectively, to find incorrect tuples for human validation, while Icarus [28] presents the expert with impactful subsets for data completion. Visual summary tools such as Profiler [184] use statistics to find data quality issues. When guiding users with visual summaries, it is important to select optimal visual encodings to reveal relevant insights or *outliers*. This can be informed from recent work by Correll et al [185], which empirically evaluated different visual encodings on their effectiveness in revealing data quality issues.

Edits

In addition to navigating data sets, amplification systems can also guide experts by suggesting data transformations to edit the data during the cleaning and preparation stage [4,28,103]. However, even in this case, transparency is required. This is evidenced by the fact that in empirical studies of Proactive Wrangler [101], users often ignored the suggested transformation but then manually performed the same one because the semantics of the operation were unclear. Methods

to aid in data transformation transparency include showing previews and transitions of the data changes [195] resulting from the transformation operation.

Interaction

Along with making system internals explainable [10], allowing experts to interact and modify data and the output of algorithms increases their trust in amplification systems [11]. For empiric antibiotic recommendation [196], this can involve allowing the health care provider to edit model features. Providing interaction comes at the cost of maintaining strict latency constraints since experts expect to see the results of interaction almost immediately [137]. Techniques for maintaining interactive performance include sampling [197] and predictive prefetching [198]. Interaction modes can include data transformation suggestions and what-if analysis.

Data Transformation

The mode of interaction for data transformation in expertise amplification systems also needs to cater to their background and training. For example, transformations should be presented as *natural language* statements [4] as opposed to code snippets [97,154]. Although graphical user interfaces can decrease trust and control for system administrators [199], they are needed in amplification systems. Gestural query systems, such as GestureDB [62] and DBTouch [200], and *direct manipulation interfaces* might be preferable to domain experts who are unfamiliar with SQL. Furthermore, domain experts' affinity for spreadsheet tools [104] motivates designing systems with spreadsheet interfaces but advanced querying capabilities such as Dataspread [201] and Sieuferd [202].

What-if Analysis

To support collaborative decision making, amplification systems should allow for what-if analysis, where domain experts can apply or test different *decisions* and *assumptions* and see how it affects the data set. Collaborative decision making is important for consensus and conflict resolution. Domain experts are highly trained and experienced individuals in their fields, which affects how they interact with systems [203,204]. Data pipeline tasks that require their input need them to apply knowledge from training and experience [28]. Such tasks inherently require judgment, which can be biased and can vary between and within domain experts [205]. To account for this bias, consensus from multiple experts is needed. However, unlike crowdworkers, where differences in results can indicate bad actors entering random choices [18,206,207], in the case of domain experts, they reveal differing judgments. As such, automatic conflict resolution [208], such as majority voting, cannot be used because disagreements require expert discussion [22]. Collaboration is required for conflict resolution, and what-if analysis can speed up this process. Capturing and sharing metadata is also useful for collaboration [209-212].

Acceleration

Time constraints of domain experts necessitate the need to accelerate their input provision. This involves designing interfaces that aid the expert's task and building interactions that interpolate from edits to generalize to multiple data points.

Interface Design

Most experts use structured interfaces such as forms [213] or free-text notes [214] for data entry or querying and spreadsheet interfaces for data exploration [104]. Following *user-centered interface* design and adhering to latency constraints is even more essential for these systems. Query interface layouts can be optimized by using statistical properties of the data [215-217] and prior query logs [64,218], while spreadsheet interfaces can be improved by incorporating higher expressibility [201,202]. The Usher [216] system, an example of the former, uses a probabilistic model on prior input form data to optimize the form structure. This involves showing highly selective data attributes at the beginning of the form to reduce the complexity at later stages, thus reducing the scope of error and accelerating input provision.

Generalize

An advantage of building systems for domain experts is that domain-specific information can be used to accelerate their input. For example, Icarus [28] uses the organism and antibiotic hierarchy encoded as foreign-key relations in the database to generalize a single edit to a rule that fills in multiple cells, accelerating the data completion process. In another example, the system by Cai et al [29] allows domain experts to refine result sets with domain-specific concepts extracted from image embeddings.

Case Study

We illustrate our taxonomy with a case study from a representative clinical data project: modeling empiric antibiotic treatment (Figure 3). We apply the 4 dimensions of amplification to the 3 stages of the pipeline. This is summarized in Table 2.

Figure 3. Data pipeline for empiric antibiotic prediction. EHR: electronic health record.

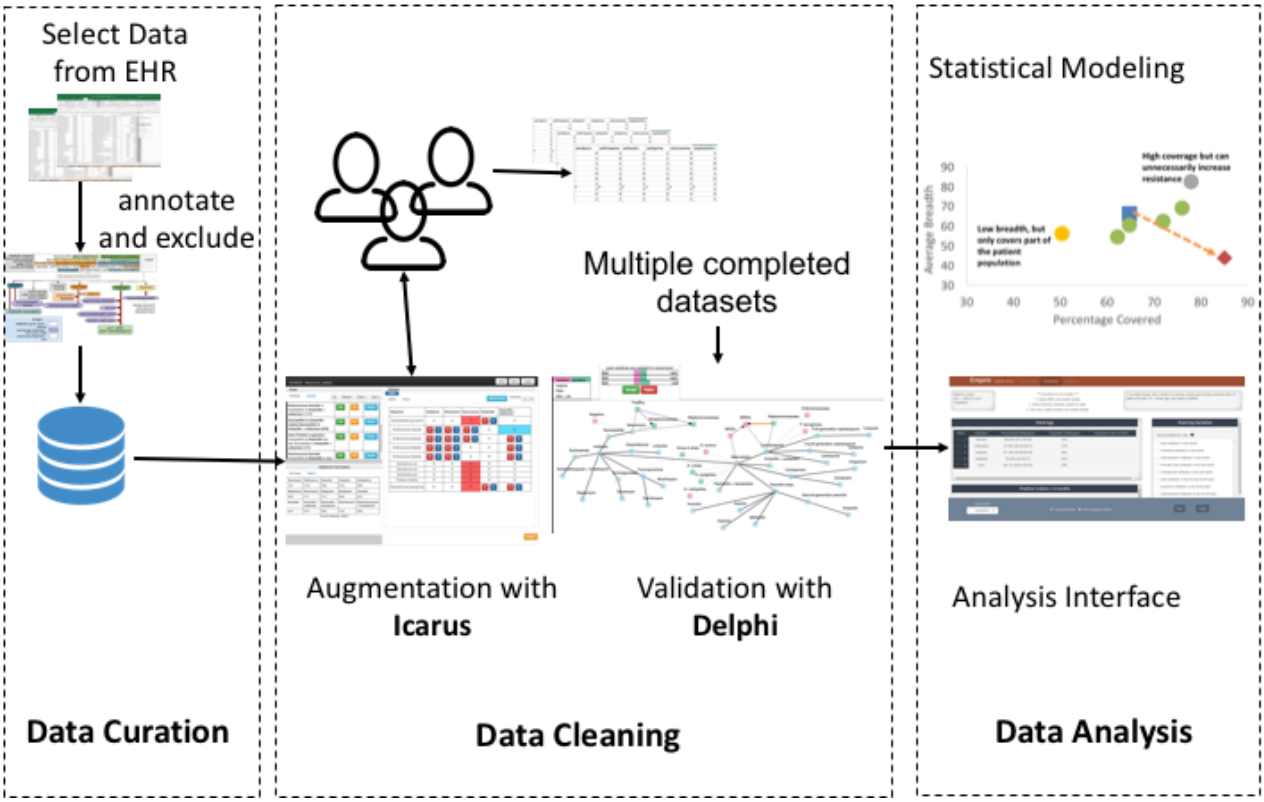


Table 2. Applying 4 dimensions of amplification to the clinical data pipeline for empiric antibiotic prediction.

Domain expert task	Amplification
Data curation	
Identify variables of interest, validate patients included in the cohort, and make domain-specific exclusionary rules	• *Summarization*: present distribution of variables of interest • *Guidance*: suggesting additional variables based on the selected ones • *Interactions*: allow expert to select and remove data points • *Acceleration*: suggest criteria based on the domain expert's inclusion and exclusion
Data cleaning	
Augmentation	
Fill in unreported microbiology susceptibilities with rules	• *Summarization*: preview a rule by showing distribution of the cells that will be impacted • *Guidance*: show high-impact data subsets for edits • *Interactions*: direct edits on interface and indirect edits via rules • *Acceleration*: suggest general rules based on the domain expert's single edit
Validation	
Validate data augmentation by examining rule set and consolidating them to remove conflicts	• *Summarization*: visual summary of rules and their relations • *Guidance*: node size guides user to high-conflict areas • *Interactions*: edit rule set by accepting and rejecting rules • *Acceleration*: automatically remove redundant rules
Data analysis	
Understand the model and its predictions for individuals and different patient subpopulations	• *Summarization*: show probability of coverage with confidence interval • *Guidance*: highlight covariates of concern • *Interactions*: allow domain expert to select covariates to include • *Acceleration*: show similar patients for who the model should be updated

At the data curation level, our domain expert, Lucy, must provide the cohort definition along with variables of interest (eg, demographics, comorbidities, allergies, etc) to a data engineer, who pulls the relevant data from the EHR data warehouse. After the data pull, Lucy looks through the initial set and formulates additional exclusion rules to ensure that it matches the clinical case definition. To implement these rules, the data engineer annotates the data with microbiology classification information of the UMLS metathesaurus [61]. This process could be improved with an expertise amplification system. The system should *summarize* data by showing the distribution of variables with linked brushing and filtering so that Lucy could see how the variable distributions are correlated. It could *guide* Lucy by suggesting correlated variables to the ones she selects. During validation of the cohort, Lucy should *interactively* be able to select data points to include. Finally, the system should be able to *accelerate* Lucy's validation by suggesting exclusion rules based on her interactions.

After the cohort is finalized, Lucy faces a data cleaning task. The microbiology laboratory provides data for only a subset of antibiotics based on domain characteristics and institutional preferences. When using these data for predictive modeling, the unreported values must be filled by domain experts. To address this, we built Icarus [28] to amplify expertise in data augmentation. Icarus *guides* the domain expert by showing them high-impact data subsets for edits. It allows both direct *interactions* via edits and indirect *interactions* via rules. Finally, Icarus *accelerates* task completion by leveraging the UMLS classification to suggest general rules based on the domain expert's single edit. It also allows the domain expert to preview the impact of a rule by *summarizing* the cells that will be impacted.

Owing to the subjective nature of this task, multiple domain experts need to come to consensus on unreported values. To amplify the consensus process, we designed Delphi [22], which visualizes the conflicts and redundancies in domain expert rules. It provides an overview of the data by visually *summarizing* the antibiotics and related rules in a node-link diagram. The node sizes *guide* the expert to regions of high conflict by encoding the number of data points affected. It allows domain experts to *interactively* edit the rule set by accepting and rejecting rules. Finally, it *accelerates* the domain experts' task completion by automatically removing redundant rules after each edit.

Once domain experts have come to a consensus, the data set is ready for analysis. Our data scientist uses penalized logistic regression to model resistance [219]. During this stage, Lucy provides insights on the different variables and their relations. After model creation, Lucy can analyze and validate the results of the interactive analysis. For a given patient, the system should *summarize* its results by showing the probability of coverage along with confidence intervals. It should *guide* Lucy by drawing attention to any abnormal covariates whose value significantly deviates from others in the cohort. It should allow Lucy to *interactively* select covariates and rerun the model for a specific patient. It should *accelerate* the analysis by showing similar patients for whom the model should also be updated.

Discussion

We have provided examples from the informatics literature to motivate the need for domain expert involvement in all steps of clinical data pipelines, from curation to analysis. Although this work is based on our experiences, we have done our best to do a targeted interdisciplinary review that can serve as a guide to clinical data projects. Our work is related to previous surveys in visual analytics in health care [188] and interactive systems [137]. Our survey is unique in that it provides a taxonomy on designing systems for amplifying expertise and focuses on the clinical data pipeline. Specifically, expertise amplification involves summarization, guidance, interactivity, and acceleration. Our case study illustrates how these can be applied to a clinical data pipeline.

Conclusions

Effectively engaging domain experts is crucial for the success of data-driven workflows. We provide a novel framework for developing systems that amplify domain expertise. Amplification systems should summarize data, guide domain experts' data navigation, allow domain experts to interact and update algorithms, and finally accelerate their task by learning from their interactions. This framework draws on research from multiple computer science disciplines. As we move toward data-driven workflows, interdisciplinary methods are necessary for the greatest impact. Empowering stakeholders to interact with the data directly can lead to faster and more impactful insights and decision making, which is vital for democratizing data to benefit society.

Acknowledgments

The research reported in this paper was supported by the National Institute of Allergy and Infectious Diseases of National Institute of Health (NIH) under R01AI116975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The National Science Foundation also supports this work under awards IIS-1422977, IIS-1527779, and CAREER IIS-1453582.

Conflicts of Interest

None declared.

References

1. Iandola F, Han S, Moskewicz M, Ashraf K, Dally W, Keutzer K. Squeezenet: alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv 2016 epub ahead of print [FREE Full text] [doi: 10.1109/CVPRW.2018.00215]
2. Ribeiro M, Singh S, Guestrin C. Why Should I Trust You: Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'16; August 13-17, 2016; San Francisco, USA. [doi: 10.1145/2939672.2939778]
3. Zarifis K, Papakonstantinou Y. ViDeTTe Interactive Notebooks. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. 2018 Presented at: HILDA'18; June 1-5, 2018; Houston, TX, USA. [doi: 10.1145/3209900.3209907]
4. Kandel S, Paepcke A, Hellerstein J, Heer J. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011 Presented at: CHI'11; March 15-17, 2011; Vancouver BC Canada p. 3363-3372. [doi: 10.1145/1978942.1979444]
5. Hu K, Orghian D, Hidalgo C. DIVE: A Mixed-Initiative System Supporting Integrated Data Exploration Workflows. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. 2018 Presented at: HILDA'18; June 1-5, 2018; Houston, TX, USA. [doi: 10.1145/3209900.3209910]
6. Kraska T. Northstar: an interactive data science system. Proc VLDB Endow 2018 Aug 1;11(12):2150-2164. [doi: 10.14778/3229863.3240493]
7. Wongsuphasawat K, Moritz D, Anand A, Mackinlay J, Howe B, Heer J. Voyager: exploratory analysis via faceted browsing of visualization recommendations. IEEE Trans Visual Comput Graphics 2016 Jan;22(1):649-658. [doi: 10.1109/tvcg.2015.2467191]
8. Xin D, Macke S, Ma L, Liu J, Song S, Parameswaran A. Helix: holistic optimization for accelerating iterative machine learning. Pro. VLDB Endow 2018 Dec 1;12(4):446-460. [doi: 10.14778/3297753.3297763]
9. Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, et al. The building blocks of interpretability. Distill 2018 Mar;3(3) [FREE Full text] [doi: 10.23915/distill.00010]
10. Yeomans M, Shah A, Mullainathan S, Kleinberg J. Making sense of recommendations. J Behav Dec Making 2019 Feb 14;32(4):403-414. [doi: 10.1002/bdm.2118]
11. Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manag Sci 2018 Mar;64(3):1155-1170. [doi: 10.1287/mnsc.2016.2643]
12. Zhang J, Wang Y, Molino P, Li L, Ebert DS. Manifold: a model-agnostic framework for interpretation and diagnosis of machine learning models. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):364-373. [doi: 10.1109/tvcg.2018.2864499]
13. Poursabzi-Sangdeh F, Goldstein D, Hofman J, Vaughan J, Wallach H. Manipulating and measuring model interpretability. arXiv preprint 2018.
14. Yin M, Vaughan J, Wallach H. Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of CHI Conference on Human Factors in Computing Systems. 2019 Presented at: Conference on Human Factors in Computing Systems; May 4-9, 2019; Scottish Event Campus Ltd, Glasgow, United Kingdom p. 1-12. [doi: 10.1145/3290605.3300509]
15. Haas D, Ansel J, Gu L, Marcus A. Argonaut: macrotask crowdsourcing for complex data processing. Proc VLDB Endow 2015 Aug;8(12):1642-1653. [doi: 10.14778/2824032.2824062]
16. Parameswaran A, Sarma AD, Garcia-Molina H, Polyzotis N, Widom J. Human-assisted graph search: it's okay to ask questions. In: Proceedings of the VLDB Endowment. 2011 Feb Presented at: International Conference on Very Large Databased; 2011; Seattle Washington p. 267-278. [doi: 10.14778/1952376.1952377]
17. Parameswaran A, Garcia-Molina H, Park H, Polyzotis N, Ramesh A, Widom J. Crowdscreen: Algorithms for filtering data with humans. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. 2012 Presented at: ACM SIGMOD/PODS conference; 2012; Scottsdale, Arizona, USA p. 361-372. [doi: 10.1145/2213836.2213878]
18. Par H, Pang R, Parameswaran A, Garcia-Molina H, Polyzotis N, Widom J. An overview of the deco system: data model and query language; query processing and optimization. ACM SIGMOD Record 2013 Jan 17;41(4):22-27. [doi: 10.1145/2430456.2430462]
19. Lasecki W, Rzeszotarski J, Marcus A, Bigham J. The effects of sequence delay on crowd work. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15: CHI Conference on Human Factors in Computing Systems; 2015; Seoul Republic of Korea p. 1375-1378. [doi: 10.1145/2702123.2702594]
20. Retelny S, Robaszkiewicz S, To A, Lasecki WS, Patel J, Rahmati N, et al. Expert Crowdsourcing With Flash Teams. In: Proceedings of the 27th annual ACM symposium on User interface software and technology. 2014 Presented at: UIST'14; October 5-8, 2014; Honolulu, Hawaii, USA. [doi: 10.1145/2642918.2647409]
21. Dror IE, Kukucka J, Kassin SM, Zapf PA. When expert decision making goes wrong: consensus, bias, the role of experts, and accuracy. J Appl Res Memory Cogn 2018 Mar;7(1):162-163. [doi: 10.1016/j.jarmac.2018.01.007]
22. Rahman P, Chen J, Hebert C, Pancholi P, Lustberg M, Stevenson K, et al. Exploratory Visualizations of Rules for Validation of Expert Decisions. In: DSIA Workshop, IEEE VIS. 2018 Presented at: Workshop on Data Systems for Interactive Analysis (DSIA); October 2018; Berlin, Germany URL: https://www.researchgate.net/publication/331177971_Exploratory_Visualizations_of_Rules_for_Validation_of_Expert_Decisions
23. Robbins R. Physicians generate an average 2.4 million a year per hospital. Southwest J Pulm Crit Care 2019;18:61 [FREE Full text]

24. Dzau VJ, Kirch DG, Nasca TJ. To care is human — collectively confronting the clinician-burnout crisis. N Engl J Med 2018 Jan 25;378(4):312-314. [doi: 10.1056/nejmp1715127]
25. Saitwal H, Feng X, Walji M, Patel V, Zhang J. Assessing performance of an electronic health record (EHR) using cognitive task analysis. Int J Med Inform 2010 Jul;79(7):501-506. [doi: 10.1016/j.ijmedinf.2010.04.001] [Medline: 20452274]
26. Heer J. Agency plus automation: designing artificial intelligence into interactive systems. Proc Natl Acad Sci U S A 2019 Feb 5;116(6):1844-1850. [doi: 10.1073/pnas.1807184115] [Medline: 30718389]
27. Krishnan S, Haas D, Franklin MJ, Wu E. Towards reliable interactive data cleaning: a user survey and recommendations. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. 2016 Presented at: SIGMOD/PODS'16: International Conference on Management of Data; June, 2016; San Francisco California. [doi: 10.1145/2939502.2939511]
28. Rahman P, Hebert C, Nandi A. ICARUS: minimizing human effort in iterative data completion. Proc VLDB Endow 2018 Sep 01;11(13):2263-2276. [doi: 10.14778/3275366.3275374]
29. Cai CJ, Reif E, Hegde N, et al. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI'19; May 4-9, 2019; Glasgow, Scotland, UK. [doi: 10.1145/3290605.3300234]
30. Clayton PD, Naus SP, Bowes WA, Madsen TS, Wilcox AB, Orsmond G, et al. Physician use of electronic medical records: issues and successes with direct data entry and physician productivity. AMIA Annu Symp Proc 2005:141-145 [FREE Full text] [Medline: 16779018]
31. Ganapathi E, Chen Y. Data Quality: Experiences and Lessons From Operationalizing Big Data. In: IEEE International Conference on Big Data. 2016 Presented at: Big Data'16; December 5-8, 2016; Washington, DC, USA. [doi: 10.1109/bigdata.2016.7840769]
32. Collins SA, Gesner E, Mar PL, Colburn DM, Rocha RA. Prioritization and refinement of clinical data elements within EHR systems. AMIA Annu Symp Proc 2016;2016:421-430 [FREE Full text] [Medline: 28269837]
33. Blaisure JC, Ceusters WM. Improving the 'Fitness for Purpose' of common data models through realism based ontology. AMIA Annu Symp Proc 2017;2017:440-447 [FREE Full text] [Medline: 29854108]
34. Pan X, Cimino J. Identifying the clinical laboratory tests from unspecified "Other Lab Test" data for secondary use. AMIA Annu Symp Proc 2015;2015:1018-1023 [FREE Full text] [Medline: 26958239]
35. Cui L, Huang Y, Tao S, Lhatoo SD, Zhang GQ. ODaCCI: ontology-guided data curation for multisite clinical research data integration in the NINDS center for SUDEP research. AMIA Annu Symp Proc 2016;2016:441-450 [FREE Full text] [Medline: 28269839]
36. Hall ES, Connolly N, Jones DE, DeFranco EA. Integrating public data sets for analysis of maternal airborne environmental exposures and stillbirth. AMIA Annu Symp Proc 2014;2014:599-605 [FREE Full text] [Medline: 25954365]
37. Gouripeddi R, Facelli JC, Bradshaw RL, Schultz D, LaSalle B, Warner PB, et al. FURTHeR: an infrastructure for clinical, translational and comparative effectiveness research. AMIA. 2013. URL: https://knowledge.amia.org/amia-55142-a2013e-1.580047/t-10-1.581994/f-010-1.581995/a-184-1.582011/ap-247-1.582014?qr=1 [accessed 2020-10-14]
38. Chen X, Wang F. Integrative spatial data analytics for public health studies of New York state. AMIA Annu Symp Proc 2016;2016:391-400 [FREE Full text] [Medline: 28269834]
39. Ying L. Combining Heterogeneous Databases to Detect Adverse Drug Reactions. Columbia University. 2015. URL: https://academiccommons.columbia.edu/doi/10.7916/D8Z60ND1 [accessed 2020-10-14]
40. Clarkson MD, Whipple ME. Variation in the representation of human anatomy within digital resources: implications for data integration. AMIA Annu Symp Proc 2018;2018:330-339 [FREE Full text] [Medline: 30815072]
41. Berrios DC, Beheshti A, Costes SV. Fairness and usability for open-access omics data systems. AMIA Annu Symp Proc 2018;2018:232-241 [FREE Full text] [Medline: 30815061]
42. Farach O, McGettrick C, Tirrell C, Evans C, Mesa A, Rozenblit L. RexMart: An Open Source Tool for Exploring and Sharing Research Data without Compromising Data Integrity. AMIA. 2014. URL: https://figshare.com/articles/RexMart_An_Open_Source_Tool_for_Exploring_and_Sharing_Research_Data_without_Compromising_Data_Integrity/1262228/1 [accessed 2020-10-14]
43. Maldonado JA, Marcos M, Fernández-Breis JT, Parcero E, Boscá D, Legaz-García MD, et al. A platform for exploration into chaining of web services for clinical data transformation and reasoning. AMIA Annu Symp Proc 2016;2016:854-863 [FREE Full text] [Medline: 28269882]
44. Zhang Y, Tang B, Jiang M, Wang J, Xu H. Domain adaptation for semantic role labeling of clinical text. J Am Med Inform Assoc 2015 Sep;22(5):967-979 [FREE Full text] [doi: 10.1093/jamia/ocu048] [Medline: 26063745]
45. Cui L, Tao S, Zhang GQ. A semantic-based approach for exploring consumer health questions using UMLS. AMIA Annu Symp Proc 2014;2014:432-441 [FREE Full text] [Medline: 25954347]
46. Dong XL, Rekatsinas T. Data Integration and Machine Learning: A Natural Synergy. In: Proceedings of the 2018 International Conference on Management of Data. 2018 Presented at: SIGMOD'18; June 1-5, 2018; Houston, Texas. [doi: 10.1145/3183713.3197387]
47. Stonebraker M, Ilyas I. Data integration: the current status and the way forward. IEEE Data Eng Bull 2018;41:3-9 [FREE Full text]

48. Fernandez C, Madden S. Termite: a system for tunneling through heterogeneous data. ArXiv 2019 epub ahead of print. [doi: 10.1145/3329859.3329877]
49. Thirumuruganathan S, Tang N, Ouzzani M. Data curation with deep learning vision: towards self-driving data curation. arXiv 2018:1384 [FREE Full text]
50. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov models. Bioinformatics 2016 Sep 15;32(18):2839-2846 [FREE Full text] [doi: 10.1093/bioinformatics/btw343] [Medline: 27283952]
51. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 2014 Feb;47:1-10 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.006] [Medline: 24393765]
52. Wei C, Kao H, Lu Z. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. Biomed Res Int 2015;2015:918710 [FREE Full text] [doi: 10.1155/2015/918710] [Medline: 26380306]
53. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 2015 Jan 19;7(S1). [doi: 10.1186/1758-2946-7-s1-s2]
54. Wei C, Harris BR, Kao H, Lu Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics 2013 Jun 1;29(11):1433-1439 [FREE Full text] [doi: 10.1093/bioinformatics/btt156] [Medline: 23564842]
55. Hielscher T, Niemann U, Preim B, Völzke H, Ittermann T, Spiliopoulou M. A framework for expert-driven subpopulation discovery and evaluation using subspace clustering for epidemiological data. Expert Syst Appl 2018 Dec;113:147-160. [doi: 10.1016/j.eswa.2018.07.003]
56. Zhang Z, Gotz D, Perer A. Iterative cohort analysis and exploration. Inf Vis 2014 Mar 19;14(4):289-307. [doi: 10.1177/1473871614526077]
57. Malik S, Du F, Monroe M, Onukwugha E, Plaisant C, Shneiderman B. An evaluation of visual analytics approaches to comparing cohorts of event sequences. 2014 Presented at: InEHRVis Workshop on Visualizing Electronic Health Record Data at VIS (Vol. 14); 2014 Nov 9; -.
58. Nargesian F, Pu Q, Zhu E, Bashardoost BG, Miller R. Optimizing organizations for navigating data lakes. arXiv 2018:7024 [FREE Full text]
59. Nargesian F, Zhu E, Pu KQ, Miller RJ. Table union search on open data. Proc VLDB Endow 2018 Mar;11(7):813-825. [doi: 10.14778/3192965.3192973]
60. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
61. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
62. Nandi A, Jiang L, Mandel M. Gestural query specification. Proc VLDB Endow 2013 Dec;7(4):289-300. [doi: 10.14778/2732240.2732247]
63. Zhang H, Raj V, Sellam T, Wu E. Precision Interfaces for Different Modalities. In: Proceedings of the 2018 International Conference on Management of Data. 2018 Presented at: SIGMOD'18; June 1-5, 2018; Houston, Texas. [doi: 10.1145/3183713.3193570]
64. Zhang H, Wu E. Mining precision interfaces from query logs. ArXiv 2019 epub ahead of print. [doi: 10.1145/3299869.3319872]
65. Peterson KJ, Jiang G, Brue SM, Liu H. Leveraging terminology services for extract-transform-load processes: a user-centered approach. AMIA Annu Symp Proc 2016;2016:1010-1019 [FREE Full text] [Medline: 28269898]
66. 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. [doi: 10.1109/jbhi.2017.2767063]
67. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable deep models for ICU outcome prediction. AMIA Annu Symp Proc 2016;2016:371-380 [FREE Full text] [Medline: 28269832]
68. Adibuzzaman M, DeLaurentis P, Hill J, Benneyworth BD. Big data in healthcare - the promises, challenges and opportunities from a research perspective: a case study with a model database. AMIA Annu Symp Proc 2017;2017:384-392 [FREE Full text] [Medline: 29854102]
69. Dallachiesa A, Ebaid, A, Eldawy, A, Elmagarmid, A, Ilyas, I, Ouzzani M, Tang N. NADEEF: a commodity data cleaning system. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013 Presented at: SIGMOD '13; 2013; New York. [doi: 10.1145/2463676.2465327]
70. Mayfield C, Neville J, Prabhakar S. Eracer: a database approach for statistical inference and data cleaning. 2010. URL: https://orion.cs.purdue.edu/docs/eracer.pdf [accessed 2020-10-14]
71. Meduri VV, Papotti P. Towards user-aware rule discovery. In: Information Search, Integration, and Personalization. Cham: Springer; 2017:3-17.
72. Wang J, Tang N. Dependable data repairing with fixing rules. J Data and Information Quality 2017 Jul 17;8(3-4):1-34. [doi: 10.1145/3041761]
73. Wang J, Krishnan S, Franklin MJ, Goldberg K, Kraska T, Milo T. A sample-and-clean framework for fast and accurate query processing on dirty data. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. 2014 Presented at: SIGMOD'14; June 2014; Snowbird Utah USA p. 469-480. [doi: 10.1145/2588555.2610505]

74. Krishnan S, Wang J, Wu E, Franklin M, Goldberg K. Activeclean: interactive data cleaning while learning convex loss models. arXiv 2016:2117-2120 [FREE Full text] [doi: 10.1145/2882903.2899409]
75. Xu J, Kalashnikov DV, Mehrotra S. Query aware determinization of uncertain objects. IEEE Trans Knowl Data Eng 2015 Jan;27(1):207-221. [doi: 10.1109/tkde.2013.170]
76. Wie Z, Link S. DataProf: Semantic Profiling for Iterative Data Cleansing and Business Rule Acquisition. In: Proceedings of the 2018 International Conference on Management of Data. 2018 Presented at: SIGMOD'18; June 1-5, 2018; Houston, Texas. [doi: 10.1145/3183713.3193544]
77. Yakout M, Elmagarmid AK, Neville J, Ouzzani M, Ilyas IF. Guided data repair. Proc VLDB Endow 2011 Feb;4(5):279-289. [doi: 10.14778/1952376.1952378]
78. Thirumuruganathan S, Berti-Equille L, Ouzzani M, Quiane-Ruiz J, Tang N. UGuide: user-guided discovery of FD-detectable errors. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-19, 2017; Chicago, USA. [doi: 10.1145/3035918.3064024]
79. Park H, Widom J. Crowdfill: collecting structured data from the crowd. ACM SIGMOD 2014 [FREE Full text] [doi: 10.1145/2588555.2610503]
80. Huang Z, Ye H. Auto-Detect: Data-Driven Error Detection in Tables. In: Proceedings of the 2018 International Conference on Management of Data. 2018 Presented at: SIGMOD'18; June 1-5, 2018; Houston, Texas. [doi: 10.1145/3183713.3196889]
81. Chu X, Morcos J, Ilyas I, Ouzzani M, Papotti P, Tang N, et al. Katara: A data cleaning system powered by knowledge bases and crowdsourcing. ACM SIGMOD 2015 [FREE Full text] [doi: 10.1145/2723372.2749431]
82. Hao S, Tang N, Li G, Li J. Cleaning Relations Using Knowledge Bases. In: IEEE 33rd International Conference on Data Engineering. 2017 Presented at: ICDE'17; April 19-22, 2017; San Diego, CA, USA. [doi: 10.1109/icde.2017.141]
83. Biessmann F, Salinas D, Schelter S, Schmidt P, Lange D. 'Deep' Learning for Missing Value Imputationin Tables with Non-Numerical Data. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018 Presented at: CIKM'18; October 22-26, 2018; Turin, Italy. [doi: 10.1145/3269206.3272005]
84. Rekatsinas T, Chu X, Ilyas IF, Ré C. HoloClean: holistic data repairs with probabilistic inference. Proc VLDB Endow 2017 Aug 1;10(11):1190-1201. [doi: 10.14778/3137628.3137631]
85. Chu X, Ilyas IF, Papotti P. Holistic Data Cleaning: Putting Violations Into Context. In: 29th International Conference on Data Engineering. 2013 Presented at: ICDE'13; April 8-12, 2013; Brisbane, QLD, Australia. [doi: 10.1109/icde.2013.6544847]
86. Schreibstein L, Newton-Dame R, McVeigh KH, Perlman SE, Singer J, Harris TG, et al. Missing data in an electronic health record-based population health surveillance system. AMIA 2014.
87. Divita G, Zeng QT, Gundlapalli AV, Duvall S, Nebeker J, Samore MH. Sophia: a expedient UMLS concept extraction annotator. AMIA Annu Symp Proc 2014:467-476. [Medline: 25954351]
88. Rumeng L, Jagannatha Abhyuday N, Hong Y. A hybrid neural network model for joint prediction of presence and period assertions of medical events in clinical notes. AMIA Annu Symp Proc 2017:1149-1158. [Medline: 29854183]
89. 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]
90. Bowles KH, Ratcliffe SJ, Naylor MD, Holmes JH, Keim SK, Flores EJ. Nurse generated EHR data supports post-acute care referral decision making: development and validation of a two-step algorithm. AMIA Annu Symp Proc 2017;2017:465-474 [FREE Full text] [Medline: 29854111]
91. Shivade C, Hebert C, Regan K, Fosler-Lussier E, Lai AM. Automatic data source identification for clinical trial eligibility criteria resolution. AMIA Annu Symp Proc 2016;2016:1149-1158 [FREE Full text] [Medline: 28269912]
92. Norman C, Leeflang M, Névéol A. Data extraction and synthesis in systematic reviews of diagnostic test accuracy: a corpus for automating and evaluating the process. AMIA Annu Symp Proc 2018;2018:817-826 [FREE Full text] [Medline: 30815124]
93. Chandar P, Yaman A, Hoxha J, He Z, Weng C. Similarity-based recommendation of new concepts to a terminology. AMIA Annu Symp Proc 2015;2015:386-395 [FREE Full text] [Medline: 26958170]
94. Kavuluru R, Rios A. Automatic assignment of non-leaf MeSH terms to biomedical articles. AMIA Annu Symp Proc 2015;2015:697-706 [FREE Full text] [Medline: 26958205]
95. Feller DJ, Zucker J, Don't Walk OB, Srikishan B, Martinez R, Evans H, et al. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. AMIA Annu Symp Proc 2018;2018:422-429 [FREE Full text] [Medline: 30815082]
96. Afshar M, Joyce C, Oakey A, Formanek P, Yang P, Churpek MM, et al. A computable phenotype for acute respiratory distress syndrome using natural language processing and machine learning. AMIA Annu Symp Proc 2018;2018:157-165 [FREE Full text] [Medline: 30815053]
97. Ratner AJ, Bach SH, Ehrenberg HR, Ré C. Snorkel: Fast Training Set Generation for Information Extraction. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-19, 2017; Chicago, USA. [doi: 10.1145/3035918.3056442]
98. Halpern Y, Choi Y, Horng S, Sontag D. Using anchors to estimate clinical state without labeled data. AMIA Annu Symp Proc 2014;2014:606-615 [FREE Full text] [Medline: 25954366]

99. Raman V, Hellerstein J. Potter's Wheel: an Interactive Framework for Data Cleaning. Univesity of Berkeley. URL: http://www/ cs.berkeley. edu/? rshankar/papers/pwheel.pdf, 2000 [accessed 2020-09-28]
100. Stolte C, Tang D, Hanrahan P. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. IEEE Trans Visual Comput Graphics 2002;8(1):52-65. [doi: 10.1109/2945.981851]
101. Guo PJ, Kandel S, Hellerstein JM, Heer J. Proactive Wrangling: Mixed-initiative End-user Programming of Data Transformation Scripts. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. 2011 Presented at: UIST'11; March 17, 2011; New York, USA. [doi: 10.1145/2047196.2047205]
102. Singh R, Gulwani S. Learning semantic string transformations from examples. Proc VLDB Endow 2012 Apr;5(8):740-751. [doi: 10.14778/2212351.2212356]
103. Jin Z, Anderson MR, Cafarella M, Jagadish HV. Foofah: transforming data by example. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-19, 2017; Chicago, USA. [doi: 10.1145/3035918.3064034]
104. Costabile MF, Fogli D, Letondal C, Mussio P, Piccinno A. Domain-expert users and their needs of software development. HCI 2003 End-User Development Session. 2003. URL: http://giove.cnuce.cnr.it/projects/EUD-NET/pdf/Costabile-et-alCameraReady.pdf [accessed 2020-10-14]
105. Hanauer DA, Hruby GW, Fort DG, Rasmussen LV, Mendonça EA, Weng C. What is asked in clinical data request forms? A multi-site thematic analysis of forms towards better data access support. AMIA Annu Symp Proc 2014;2014:616-625 [FREE Full text] [Medline: 25954367]
106. Hruby GW, Hoxha J, Ravichandran PC, Mendonça EA, Hanauer DA, Weng C. A data-driven concept schema for defining clinical research data needs. Int J Med Inform 2016 Jul;91:1-9 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.03.008] [Medline: 27185504]
107. 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]
108. Hu Z, Melton GB, Moeller ND, Arsoniadis EG, Wang Y, Kwaan MR, et al. Accelerating chart review using automated methods on electronic health record data for postoperative complications. AMIA Annu Symp Proc 2016;2016:1822-1831 [FREE Full text] [Medline: 28269941]
109. Major V, Tanna MS, Jones S, Aphinyanaphongs Y. Reusable filtering functions for application in ICU data: a case study. AMIA Annu Symp Proc 2016;2016:844-853 [FREE Full text] [Medline: 28269881]
110. Zhao L. Controlling False Discoveries During Interactive Data Exploration. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-17, 2017; Chicago, USA. [doi: 10.1145/3035918.3064019]
111. Romero-Brufau S, Kostandy P, Maass KL, Wutthisirisart P, Sir M, Bartholmai B, et al. Development of data integration and visualization tools for the Department of Radiology to display operational and strategic metrics. AMIA Annu Symp Proc 2018;2018:942-951 [FREE Full text] [Medline: 30815137]
112. Pore M, Sengeh DM, Mugambi P, Purswani NV, Sesay T, Arnold AL, et al. Design and evaluation of a web-based decision support tool for district-level disease surveillance in a low-resource setting. AMIA Annu Symp Proc 2017;2017:1401-1410 [FREE Full text] [Medline: 29854209]
113. Iyer G, DuttaDuwarah S, Sharma A. DataScope: Interactive Visual Exploratory Dashboards for Large Multidimensional Data. In: Workshop on Visual Analytics in Healthcare. 2017 Presented at: VAHC'17; October 1, 2017; Phoenix, AZ, USA. [doi: 10.1109/vahc.2017.8387496]
114. Anscombe FJ. Graphs in statistical analysis. Am Stat 1973 Feb;27(1):17. [doi: 10.2307/2682899]
115. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings 1996 IEEE Symposium on Visual Languages. 2003 Presented at: IEEE Symposium on Visual Languages; August 6, 2002; Boulder, CO, USA. [doi: 10.1016/b978-155860915-0/50046-9]
116. Vartak M, Madden S, Parameswaran A, Polyzotis N. SeeDB: automatically generating query visualizations. Proc VLDB Endow 2014 Aug;7(13):1581-1584. [doi: 10.14778/2733004.2733035]
117. Moritz D, Wang C, Nelson GL, Lin H, Smith AM, Howe B, et al. Formalizing visualization design knowledge as constraints: actionable and extensible models in draco. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):438-448. [doi: 10.1109/tvcg.2018.2865240]
118. Demiralp A, Haas PJ, Parthasarathy S, Pedapati T. Foresight: recommending visual insights. Proc VLDB Endow 2017 Aug 1;10(12):1937-1940. [doi: 10.14778/3137765.3137813]
119. Caballero GC. Visual Analytics for Evaluating Clinical Pathways. In: Workshop on Visual Analytics in Healthcare. 2017 Presented at: VAHC'17; October 1, 2017; Phoenix, AZ, USA. [doi: 10.1109/vahc.2017.8387499]
120. Widanagamaachchi W, Livnat Y, Bremer PT, Duvall S, Pascucci V. Interactive visualization and exploration of patient progression in a hospital setting. AMIA Annu Symp Proc 2017;2017:1773-1782. [Medline: 29854248]
121. Li X, Cui L, Tao S, Zeng N, Zhang GQ. SpindleSphere: a web-based platform for large-scale sleep spindle analysis and visualization. AMIA Annu Symp Proc 2017;2017:1159-1168 [FREE Full text] [Medline: 29854184]
122. Mortensen JM, Musen MA, Noy NF. An empirically derived taxonomy of errors in SNOMED CT. AMIA Annu Symp Proc 2014;2014:899-906 [FREE Full text] [Medline: 25954397]

123. Chen ES, Melton GB, Wasserman RC, Rosenau PT, Howard DB, Sarkar IN. Mining and visualizing family history associations in the electronic health record: a case study for Pediatric Asthma. AMIA Annu Symp Proc 2015;2015:396-405 [FREE Full text] [Medline: 26958171]
124. Sockolow PS, Yang Y, Bass EJ, Bowles KH, Holmberg A, Sheryl P. Data visualization of home care admission nurses' decision-making. AMIA Annu Symp Proc 2017;2017:1597-1606 [FREE Full text] [Medline: 29854230]
125. Kummerfeld E, Anker JA, Rix A, Kushner MG. Methodological advances in the study of hidden variables: a demonstration on clinical alcohol use disorder data. AMIA Annu Symp Proc 2018;2018:710-719 [FREE Full text] [Medline: 30815113]
126. Sarawagi S, Agrawal R, Megiddo N. Discovery-driven Exploration of OLAP Data Cubes. In: International Conference on Extending Database Technology. 1998 Presented at: EDBT'98; March 23-27, 1998; Valencia, Spain. [doi: 10.1007/bfb0100984]
127. Sordo M, Tokachichu P, Vitale CJ, Maviglia SM, Rocha RA. Modeling contextual knowledge for clinical decision support. AMIA Annu Symp Proc 2017;2017:1617-1624 [FREE Full text] [Medline: 29854232]
128. Gangadhar S, Nguyen N, Pesuit JW, Bogdanov AN, Kallenbach L, Ken J, et al. Effectiveness of a cloud-based EHR clinical decision support program for body mass index (BMI) screening and follow-up. AMIA Annu Symp Proc 2017;2017:742-749 [FREE Full text] [Medline: 29854140]
129. Souissi SB, Abed M, Elhiki L, Fortemps P, Pirlot M. Reducing the toxicity risk in antibiotic prescriptions by combining ontologies with a multiple criteria decision model. AMIA Annu Symp Proc 2017;2017:1625-1634 [FREE Full text] [Medline: 29854233]
130. Cardoso SD, Chantal RD, Da Silveira M, Pruski C. Combining rules, background knowledge and change patterns to maintain semantic annotations. AMIA Annu Symp Proc 2017;2017:505-514 [FREE Full text] [Medline: 29854115]
131. Hedda M, Malin BA, Yan C, Fabbri D. Evaluating the effectiveness of auditing rules for electronic health record systems. AMIA Annu Symp Proc 2017;2017:866-875 [FREE Full text] [Medline: 29854153]
132. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD'15; August 10-13, 2015; Sydney, Australia. [doi: 10.1145/2783258.2788613]
133. Che Z, St Sauver J, Liu H, Liu Y. Deep learning solutions for classifying patients on opioid use. AMIA Annu Symp Proc 2017;2017:525-534 [FREE Full text] [Medline: 29854117]
134. Ge W, Huh JW, Park YR, Lee JH, Kim YH, Turchin A. An interpretable ICU mortality prediction model based on logistic regression and recurrent neural networks with LSTM units. AMIA Annu Symp Proc 2018;2018:460-469 [FREE Full text] [Medline: 30815086]
135. Ho KC, Speier W, El-Saden S, Arnold CW. Classifying acute ischemic stroke onset time using deep imaging features. AMIA Annu Symp Proc 2017;2017:892-901 [FREE Full text] [Medline: 29854156]
136. Ming Y, Qu H, Bertini E. RuleMatrix: visualizing and understanding classifiers with rules. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):342-352. [doi: 10.1109/tvcg.2018.2864812]
137. Rahman P, Jiang L, Nandi A. Evaluating interactive data systems. VLDB J 2019 Nov 13;29(1):119-146. [doi: 10.1007/s00778-019-00589-2]
138. Doan A, Domingos P, Halevy AY. Reconciling schemas of disparate data sources. SIGMOD Rec 2001 Jun;30(2):509-520. [doi: 10.1145/376284.375731]
139. Do HH, Rahm E. COMA - a system for flexible combination of schema matching approaches. In: Proceedings of 28th International Conference on Very Large Data Bases. 2002 Presented at: 28th International Conference on Very Large Data Bases; August 20-23, 2002; Hong Kong, China. [doi: 10.1016/b978-155860869-6/50060-3]
140. Chen K, Kannan A, Madhavan J, Halevy A. Exploring schema repositories with schemr. SIGMOD Rec 2011 Jul 18;40(1):11-16. [doi: 10.1145/2007206.2007210]
141. Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. VLDB J . 334? 2001;10(4):350. [doi: 10.1007/s007780100057]
142. Nandi A, Bernstein PA. HAMSTER: using search clicklogs for schema and taxonomy matching. Proc VLDB Endow 2009 Aug 1;2(1):181-192. [doi: 10.14778/1687627.1687649]
143. Wang Y. Synthesizing Mapping Relationships Using Table Corpus. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-17, 2017; Chicago, USA. [doi: 10.1145/3035918.3064010]
144. Ives Z, Knoblock CA, Minton S, Jacob M, Talukdar PP, Tuchinda R, et al. arXiv. 2009. URL: http://talukdar.net/papers/cidr.pdf [accessed 2020-10-14]
145. Wang J, Kraska T, Franklin MJ, Feng J. CrowdER: crowdsourcing entity resolution. Proc VLDB Endow 2012 Jul;5(11):1483-1494. [doi: 10.14778/2350229.2350263]
146. Cafarella M, Halevy A, Lee H, Madhavan J, Yu C, Wang DZ, et al. Ten years of webtables. Proc VLDB Endow 2018 Aug 1;11(12):2140-2149. [doi: 10.14778/3229863.3240492]
147. Dong XL, Halevy A, Yu C. Data integration with uncertainty. VLDB J 2008 Nov 14;18(2):469-500. [doi: 10.1007/s00778-008-0119-9]

148. Zhang CJ, Chen L, Jagadish HV, Zhang M, Tong Y. Reducing uncertainty of schema matching via crowdsourcing with accuracy rates. IEEE Trans. Knowl Data Eng 2020 Jan 1;32(1):135-151. [doi: 10.1109/tkde.2018.2881185]
149. Cafarella MJ, Halevy A, Khoussainova N. Data integration for the relational web. Proc VLDB Endow 2009 Aug 1;2(1):1090-1101. [doi: 10.14778/1687627.1687750]
150. Madhavan J, Jeffery SR, Cohen S, Dong X, Ko D, Yu C, et al. Web-scale Data Integration: You can only afford to Pay As You Go. MIT. 2007. URL: http://web.mit.edu/tibbetts/Public/CIDR_2007_Proceedings/papers/cidr07p40.pdf [accessed 2020-10-14]
151. Limaye G, Sarawagi S, Chakrabarti S. Annotating and searching web tables using entities, types and relationships. Proc VLDB Endow 2010 Sep;3(1-2):1338-1347. [doi: 10.14778/1920841.1921005]
152. Miller RJ. Open data integration. Proc VLDB Endow 2018 Aug 1;11(12):2130-2139. [doi: 10.14778/3229863.3240491]
153. Fernandez RC, Mansour E, Qahtan AA, Elmagarmid A, Ilyas I, Madden S, et al. Seeping semantics: Linking datasets using word embeddings for data discovery. 2018 Presented at: IEEE 34th International Conference on Data Engineering; October 25, 2018; Paris, France. [doi: 10.1109/icde.2018.00093]
154. He J, Veltri E, Santoro D, Li G, Mecca G, Papotti P, et al. Interactive and Deterministic Data Cleaning. In: Proceedings of the International Conference on Management of Data. 2016 Presented at: SIGMOD '16; June 2016; New York. [doi: 10.1145/2882903.2915242]
155. Bergman M, Milo T, Novgorodov S, Tan WC. Query-oriented data cleaning with oracles. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 2015 Presented at: SIGMOD '15; May, 2015; New York p. 1199-1214. [doi: 10.1145/2723372.2737786]
156. Assadi A, Milo T, Novgorodov S. Cleaning Data with Constraints and Experts. In: Proceedings of the 21st International Workshop on the Web and Databases. 2018 Presented at: WebDB'18; June 10, 2018; Houston, TX, USA. [doi: 10.1145/3201463.3201464]
157. Fan W, Geerts F, Lakshmanan L, Xiong M. Discovering conditional functional dependencies. In: Proceedings of the 25th International Conference on Data Engineering. 2011 Presented at: International Conference on Data Engineering; March 29, 2009 - April 2, 2009; Shanghai, China. [doi: 10.1109/icde.2009.208]
158. Chiang F, Miller RJ. Discovering data quality rules. Proc VLDB Endow 2008 Aug 1;1(1):1166-1177. [doi: 10.14778/1453856.1453980]
159. Golab L, Karloff H, Korn F, Srivastava D, Yu B. On generating near-optimal tableaux for conditional functional dependencies. Proc VLDB Endow 2008 Aug 1;1(1):376-390. [doi: 10.14778/1453856.1453900]
160. Cong G, Fan W, Geerts F, Jia X, Ma S. Improving data quality: consistency and accuracy. VLDB J. 2007. URL: http://homepages.inf.ed.ac.uk/wenfei/papers/vldb07-b.pdf [accessed 2020-10-14]
161. Fan W, Geerts F. Foundations of data quality management. Synth Lect Data Manag 2012 Jul 31;4(5):1-217. [doi: 10.2200/s00439ed1v01y201207dtm030]
162. Wang DZ, Dong XL, Sarma AD, Franklin MJ. Functional dependency generation and applications in pay-as-you-go data integration systems. 2009 Presented at: 12th International Workshop on the Web and Databases; June 28, 2009; Providence, Rhode Island, USA.
163. Rammelaere J, Geerts F. Explaining repaired data with CFDs. Proc VLDB Endow 2018 Jul 1;11(11):1387-1399. [doi: 10.14778/3236187.3236193]
164. Ilyas IF, Markl V, Haas P, Brown P, Aboulnaga A. CORDS: Automatic Discovery of Correlations and Soft Functional Dependencies. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data. 2004 Presented at: SIGMOD'14; June 13-18, 2004; Paris. [doi: 10.1145/1007568.1007641]
165. Asghar A, Ghenai A. Automatic discovery of functional dependencies and conditional functional dependencies: a comparative study. University of Waterloo. 2015. URL: https://cs.uwaterloo.ca/~nasghar/848.pdf [accessed 2020-10-14]
166. De Sa C, Ratner A, Ré C, Shin J, Wang F, Wu S, et al. DeepDive: declarative knowledge base construction. SIGMOD Rec 2016 Jun 2;45(1):60-67. [doi: 10.1145/2949741.2949756]
167. Varma P, Ré C. Snuba: automating weak supervision to label training data. Proc VLDB Endow 2018 Nov 1;12(3):223-236 [FREE Full text] [doi: 10.14778/3291264.3291268] [Medline: 31777681]
168. Felix C, Dasgupta A, Bertini E. The Exploratory Labeling Assistant: Mixed-Initiative Label Curation with Large Document Collections Share on. In: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology. 2018 Presented at: UIST'18; June 1-5, 2018; New York, USA. [doi: 10.1145/3242587.3242596]
169. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013 Jul;41(Web Server issue):W518-W522 [FREE Full text] [doi: 10.1093/nar/gkt441] [Medline: 23703206]
170. Kwon D, Kim S, Wei CH, Leaman R, Lu Z. ezTag: tagging biomedical concepts via interactive learning. Nucleic Acids Res 2018 Jul 2;46(W1):W523-W529 [FREE Full text] [doi: 10.1093/nar/gky428] [Medline: 29788413]
171. Nazi A, Ding B, Narasayya V, Chaudhuri S. Efficient estimation of inclusion coefficient using hyperloglog sketches. Proc VLDB Endow 2018 Jun 1;11(10):1097-1109. [doi: 10.14778/3231751.3231759]
172. Yuan X, Cai X, Yu M, Wang C, Zhang Y, Wen Y. Efficient Foreign Key Discovery Based on Nearest Neighbor Search. In: International Conference on Web-Age Information Management. 2015 Presented at: WAIM'15; June 8-10, 2015; Qingdao, China. [doi: 10.1007/978-3-319-21042-1_37]

173. Motl J, Kordik P. Foreign key constraint identification in relational databases. Czech Technical University in Prague. 2017. URL: http://ceur-ws.org/Vol-1885/106.pdf [accessed 2020-10-14]
174. Koehler H, Link S. Inclusion Dependencies Reloaded. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015 Presented at: CIKM '15; October, 2015; Melbourne Australia.
175. Chen Z, Narasayya V, Chaudhuri S. Fast foreign-key detection in Microsoft SQL server PowerPivot for Excel. Proc VLDB Endow 2014 Aug;7(13):1417-1428. [doi: 10.14778/2733004.2733014]
176. Moritz D, Howe B, Heer J. Falcon: Balancing interactive latency and resolution sensitivity for scalable linked visualizations,? University of Washington. 2019. URL: https://idl.cs.washington.edu/files/2019-Falcon-CHI.pdf [accessed 2020-10-14]
177. Kamat N, Jayachandran P, Tunga K, Nandi A. Distributed and Interactive Cube Exploration. In: 30th International Conference on Data Engineering. 2014 Presented at: ICDE'14; March 31-April 4, 2014; Chicago, IL, USA. [doi: 10.1109/icde.2014.6816674]
178. Lins L, Klosowski JT, Scheidegger C. Nanocubes for real-time exploration of spatiotemporal datasets. IEEE Trans Visual Comput Graphics 2013 Dec;19(12):2456-2465. [doi: 10.1109/tvcg.2013.179]
179. Pahins CA, Stephens SA, Scheidegger C, Comba JL. Hashedcubes: simple, low memory, real-time visual exploration of big data. IEEE Trans Visual Comput Graphics 2017 Jan;23(1):671-680. [doi: 10.1109/tvcg.2016.2598624]
180. Joglekar M, Garcia-Molina H, Parameswaran A. Interactive Data Exploration With Smart Drill-down. In: 32nd International Conference on Data Engineering. 2016 Presented at: ICDE'16; May 16-20, 2016; Helsinki, Finland. [doi: 10.1109/icde.2016.7498300]
181. Dimitriadou K, Papaemmanouil O, Diao Y. AIDE: an active learning-based approach for interactive data exploration. IEEE Trans Knowl Data Eng 2016 Nov 1;28(11):2842-2856. [doi: 10.1109/tkde.2016.2599168]
182. Psallidas F, Wu E. Smoke. Proc VLDB Endow 2018 Feb 1;11(6):719-732. [doi: 10.14778/3199517.3199522]
183. Wu E, Madden S. Scorpion. Proc VLDB Endow 2013 Jun;6(8):553-564. [doi: 10.14778/2536354.2536356]
184. Kandel S, Parikh R, Paepcke A, Hellerstein J, Heer J. Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. Stanford Univeristy. 2012. URL: http://vis.stanford.edu/papers/profiler [accessed 2020-10-14]
185. Correll M, Li M, Kindlmann G, Scheidegger C. Looks good to me: visualizations as sanity checks. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):830-839. [doi: 10.1109/tvcg.2018.2864907]
186. Wongsuphasawat K, Moritz D, Qu Z, Chang R, Ouk F, Anand A, et al. Voyager 2: Augmenting Visual Analysis with Partial View Specifications. University of Washington. 2017. URL: https://idl.cs.washington.edu/files/2017-Voyager2-CHI.pdf [accessed 2020-10-14]
187. Willett W, Heer J, Agrawala M. Scented widgets: improving navigation cues with embedded visualizations. IEEE Trans Visual Comput Graphics 2007 Nov;13(6):1129-1136. [doi: 10.1109/tvcg.2007.70589]
188. Preim B, Lawonn K. A survey of visual analytics for public health. Computer Graphics Forum 2019 Nov 28;39(1):543-580. [doi: 10.1111/cgf.13891]
189. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv 2017:9923.
190. Montavon G, Samek W, Müller K. Methods for interpreting and understanding deep neural networks. Digi Signal Process 2018 Feb;73:1-15. [doi: 10.1016/j.dsp.2017.10.011]
191. Alspaugh S, Zokaei N, Liu A, Jin C, Hearst MA. Futzing and moseying: interviews with professional data analysts on exploration practices. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):22-31. [doi: 10.1109/tvcg.2018.2865040]
192. Stoyanovich J, Howe B, Jagadish H, Miklau G. Panel: a debate on data and algorithmic ethics. Proc VLDB Endow 2018 Aug 1;11(12):2165-2167. [doi: 10.14778/3229863.3240494]
193. Cai CJ, Jongejaan J, Holbrook J. The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019 Presented at: IUI '19; March, 2019; New York.
194. Liu D, Xu P, Ren L. Tpflow: progressive partition and multidimensional pattern extraction for large-scale spatio-temporal data analysis. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):1-11. [doi: 10.1109/tvcg.2018.2865018]
195. Khan M, Xu L, Nandi A, Hellerstein JM. Data tweening: incremental visualization of data transforms. Proc VLDB Endow 2017 Feb 1;10(6):661-672.
196. Shah N, Acree ME, Patros C, Suseno M, Grant J, Fleming G, et al. A novel inpatient Antibiotic Stewardship Assistance Program (ASAP) using real-time electronic health record data, prediction modeling and epidemiologic data to provide personalized empiric antibiotic recommendations. Open Forum Infect Dis 2018 Nov;5(Suppl 1). [doi: 10.1093/ofid/ofy210.201]
197. Kamat N, Nandi A. A session-based approach to fast-but-approximate interactive data cube exploration. ACM Trans Knowl Discov Data 2018 Feb 23;12(1):1-26. [doi: 10.1145/3070648]
198. Battle L, Chang R, Stonebraker M. Dynamic prefetching of data tiles for interactive visualization. In: Proceedings of the International Conference on Management of Data. 2016 Presented at: SIGMOD '16; June, 2016; New York.
199. Takayama L, Kandogan E. Trust as an underlying factor of system administrator interface choice. In: Extended Abstracts on Human Factors in Computing Systems. 2006 Presented at: CHI EA '06; April, 2006; Montréal Québec Canada. [doi: 10.1145/1125451.1125708]

200. Idreos S, Liarou E. dbTouch: Analytics at your Fingertips. Stanford Univeristy. 2013. URL: http://www-cs-students.stanford.edu/~adityagp/courses/cs598-old/papers/dbtouch.pdf [accessed 2020-10-14]
201. Bendre M, Sun B, Zhang D, Zhou X, Chang KC, Parameswaran A. DataSpread. Proc VLDB Endow 2015 Aug;8(12):2000-2003. [doi: 10.14778/2824032.2824121]
202. Ozcan F, Koutrika G. Expressive Query Construction through Direct Manipulation of Nested Relational Results. In: Proceedings of the 2016 International Conference on Management of Data. 2016 Presented at: SIGMOD '16; June 2016; San Francisco, California.
203. Schaffer J, O'Donovan J, Michaelis J, Raglin A, Höllerer T. I can do better than your AI: expertise and explanations. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019 Presented at: IUI '19; March 2019; Marina del Ray, California. [doi: 10.1145/3301275.3302308]
204. Arnold V, Clark N, Collier PA, Leech SA, Sutton S. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. MIS Q 2006;30(1):79. [doi: 10.2307/25148718]
205. Dror IE. A hierarchy of expert performance. J Appl Res Memory Cogn 2016 Jun;5(2):121-127. [doi: 10.1016/j.jarmac.2016.03.001]
206. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods 2012 Mar;44(1):23. [doi: 10.3758/s13428-011-0124-6] [Medline: 21717266]
207. Czerwinski M, Lund A. Crowdsourcing user studies with Mechanical Turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2008 Presented at: CHI '08; April, 2008; Florence, Italy.
208. Sarma AD, Parameswaran A, Widom J. Towards Globally Optimal Crowdsourcing Quality Management: The Uniform Worker Setting. In: Proceedings of the 2016 International Conference on Management of Data. 2016 Presented at: SIGMOD '16; June, 2016; San Francisco, California.
209. Kandogan E, Roth M, Shwarz P, Hui J, Terizzano I, Christodoulakis C, et al. LabBook: Metadata-driven social collaborative data analysis. 2015 Presented at: 2015 IEEE International Conference on Big Data (Big Data); 2015; Santa Clara, California.
210. Hellerstein M, Sreekanti V, Gonzalez JE, Dalton J, Dey A, Nag S, et al. Ground: A Data Context Service. Conference on Innovative Data Systems Research. 2017. URL: http://cidrdb.org/cidr2017/papers/p111-hellerstein-cidr17.pdf [accessed 2020-10-14]
211. Kandel S, Paepcke A, Hellerstein JM, Heer J. Enterprise data analysis and visualization: an interview study. IEEE Trans Visual Comput Graphics 2012 Dec;18(12):2917-2926. [doi: 10.1109/tvcg.2012.219]
212. Kandogan E, Balakrishnan A, Haber EM, Pierce JS. From data to insight: work practices of analysts in the enterprise. IEEE Comput Grap Appl 2014 Sep;34(5):42-50. [doi: 10.1109/mcg.2014.62]
213. Jagdish HV, Nandi A, Qian L. Organic databases. In: Databases in Networked Information Systems. Berlin, Heidelberg: Springer; 2011:49-63.
214. Embi PJ, Yackel TR, Logan JR, Bowen JL, Cooney TG, Gorman PA. Impacts of computerized physician documentation in a teaching hospital: perceptions of faculty and resident physicians. J Am Med Inf Assoc 2004 Apr 2;11(4):300-309. [doi: 10.1197/jamia.m1525]
215. Rahman P, Nandi A. Transformer: a database-driven approach to generating forms for constrained interaction. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019 Presented at: IUI '19; 2019; Marina del Ray, California.
216. Chen K, Chen H, Conway N, Hellerstein JM, Parikh TS. Usher: improving data quality with dynamic forms. IEEE Trans Knowl Data Eng 2011 Aug;23(8):1138-1153. [doi: 10.1109/tkde.2011.31]
217. Gajos K, Weld DS. SUPPLE: Automatically Generating User Interfaces. Harvard University. 2004. URL: https://www.eecs.harvard.edu/~kgajos/papers/2004/supple-iui04.pdf [accessed 2020-10-14]
218. Jayapandian N, Jagadish HV. Automated creation of a forms-based database query interface. Proceedings VLDB Endowment 2008 Aug;1(1):695-709.
219. Hebert C, Gao Y, Rahman P, Dewart C, Lustberg M, Pancholi P, et al. Prediction of antibiotic susceptibility for urinary tract infection in a hospital setting. Antimicrob Agents Chemother 2020 Jun 23;64(7):02236-19. [doi: 10.1128/aac.02236-19]

Abbreviations

CDS: clinical decision support
EHR: electronic health record
HCI: human-computer interaction
NIH: National Institute of Health
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 24.04.20; peer-reviewed by M Afzal, A Benis, M Spiliopoulou; comments to author 25.05.20; revised version received 07.07.20; accepted 22.07.20; published 05.11.20.

Please cite as:
Rahman P, Nandi A, Hebert C
Amplifying Domain Expertise in Clinical Data Pipelines
JMIR Med Inform 2020;8(11):e19612
URL: https://medinform.jmir.org/2020/11/e19612
doi:10.2196/19612
PMID:33151150

Original Paper

Identifying Ectopic Pregnancy in a Large Integrated Health Care Delivery System: Algorithm Validation

Darios Getahun[1,2], MD, MPH, PhD; Jiaxiao M Shi[1], PhD; Malini Chandra[3], MS, MBA; Michael J Fassett[4], MD; Stacey Alexeeff[3], PhD; Theresa M Im[1], MPH; Vicki Y Chiu[1], MS; Mary Anne Armstrong[3], MA; Fagen Xie[1], PhD; Julie Stern[1], MPH; Harpreet S Takhar[1], MPH; Alex Asiimwe[5], PhD; Tina Raine-Bennett[2,3], MD, MPH

[1]Department of Research & Evaluation, Kaiser Permanente Southern California, Pasadena, CA, United States

[2]Department of Health Systems Science, Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, United States

[3]Division of Research, Kaiser Permanente Northern California, Oakland, CA, United States

[4]Department of Obstetrics and Gynecology, Kaiser Permanente West Los Angeles Medical Center, Los Angeles, CA, United States

[5]Bayer AG, Berlin, Germany

Corresponding Author:
Darios Getahun, MD, MPH, PhD
Department of Research & Evaluation
Kaiser Permanente Southern California
100 S. Los Robles
Pasadena, CA, 91101
United States
Phone: 1 626 564 5658
Email: Darios.T.Getahun@kp.org

Abstract

Background: Surveillance of ectopic pregnancy (EP) using electronic databases is important. To our knowledge, no published study has assessed the validity of EP case ascertainment using electronic health records.

Objective: We aimed to assess the validity of an enhanced version of a previously validated algorithm, which used a combination of encounters with EP-related diagnostic/procedure codes and methotrexate injections.

Methods: Medical records of 500 women aged 15-44 years with membership at Kaiser Permanente Southern and Northern California between 2009 and 2018 and a potential EP were randomly selected for chart review, and true cases were identified. The enhanced algorithm included diagnostic/procedure codes from the International Classification of Diseases, Tenth Revision, used telephone appointment visits, and excluded cases with only abdominal EP diagnosis codes. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and overall performance (Youden index and F-score) of the algorithm were evaluated and compared to the validated algorithm.

Results: There were 334 true positive and 166 true negative EP cases with available records. True positive and true negative EP cases did not differ significantly according to maternal age, race/ethnicity, and smoking status. EP cases with only one encounter and non-tubal EPs were more likely to be misclassified. The sensitivity, specificity, PPV, and NPV of the enhanced algorithm for EP were 97.6%, 84.9%, 92.9%, and 94.6%, respectively. The Youden index and F-score were 82.5% and 95.2%, respectively. The sensitivity and NPV were lower for the previously published algorithm at 94.3% and 88.1%, respectively. The sensitivity of surgical procedure codes from electronic chart abstraction to correctly identify surgical management was 91.9%. The overall accuracy, defined as the percentage of EP cases with correct management (surgical, medical, and unclassified) identified by electronic chart abstraction, was 92.3%.

Conclusions: The performance of the enhanced algorithm for EP case ascertainment in integrated health care databases is adequate to allow for use in future epidemiological studies. Use of this algorithm will likely result in better capture of true EP cases than the previously validated algorithm.

(JMIR Med Inform 2020;8(11):e18559) doi:10.2196/18559

KEYWORDS

ectopic pregnancy; pregnancy; validation; predictive value; electronic health records; electronic database

Introduction

Use of claims, administrative databases, and electronic health records (EHRs) allows for efficient identification of individuals with medical conditions [1]. National hospital databases and discharge diagnoses have been used extensively to monitor serious medical conditions leading to significant morbidity such as acute myocardial infarction; however, hospital databases are not sufficient in capturing serious conditions that do not necessarily require hospitalization. Ectopic pregnancy (EP), the implantation of a fertilized ovum outside of the endometrial cavity, is a serious condition that can be life threatening; however, a significant proportion of patients can be managed in the outpatient setting. Trends in EP are difficult to examine because women with EPs are increasingly managed in the outpatient setting, either medically with methotrexate injection(s) or surgically with laparoscopy [2,3]. Furthermore, women with potential EPs may be evaluated over the course of several days and medical encounters prior to the establishment of a definitive diagnosis of EP or viable or nonviable intrauterine pregnancy, making identification of true cases difficult.

Researchers have typically relied on clinical diagnosis and procedure codes extracted from outpatient care and hospital discharge databases to describe trends in EP. However, the accuracy of EP case ascertainment and the validity of study findings depend on the types of data sources and completeness of EP case ascertainment approaches. One methodology for EP case ascertainment was validated in a study by Scholes et al in 2011 [4], using claims and administrative data extracted from a large health care maintenance organization database prior to the use of EHRs and codes from the International Classification of Diseases, Tenth Revision (ICD-10). Although the sensitivity of the algorithm for capturing EP cases was higher than that of the use of standard codes, the algorithm is inherently limited by the time frame of the study, the completeness of the data, and the ability to review patients' medical information in an electronic database for true case ascertainment [5-8].

The widespread adoption of EHRs in the United States presents an opportunity to improve patient care [9,10] and provides researchers unparalleled possibilities to conduct high-quality clinical and pharmacoepidemiologic research [11,12]. EHRs provide access to more reliable and comprehensive patient health information. They are also easily transferable to other EHR systems and more cost-efficient than paper-based data sources [13-15]. Over the last decade, there have been a number of studies that evaluated the accuracy of health data (hospital discharge data, outpatient encounter data, and claims data) extracted from the EHRs of various regions of the Kaiser Permanente health care system [16-19] and other health care systems [20,21]. Published validation studies investigated demographic characteristics [17], body weight and height data [22], perinatal outcomes [18,23], phenotype for genomic study [21], and phenotype of HIV infection [20]. However, to our knowledge, there is no study that has assessed the validity of EP case ascertainment using EHRs for validation and the potential impact of changes in the data over time (pre-EHR vs EHR era). There is substantial practice pattern variation over time, across institutions and health care providers. The Scholes et al algorithm was developed 10 years ago at two institutions with potentially different practice environments than the setting of this study. Furthermore, the data for the Scholes et al algorithm came largely from contracting hospitals for inpatient care, which may have disparate practice and coding patterns. Therefore, validating the algorithm in a different time frame and setting is necessary to conduct future studies describing the temporal trends of EP incidence and treatment modalities. This study aimed to develop an enhanced algorithm that builds on the previously validated algorithm [4].

Methods

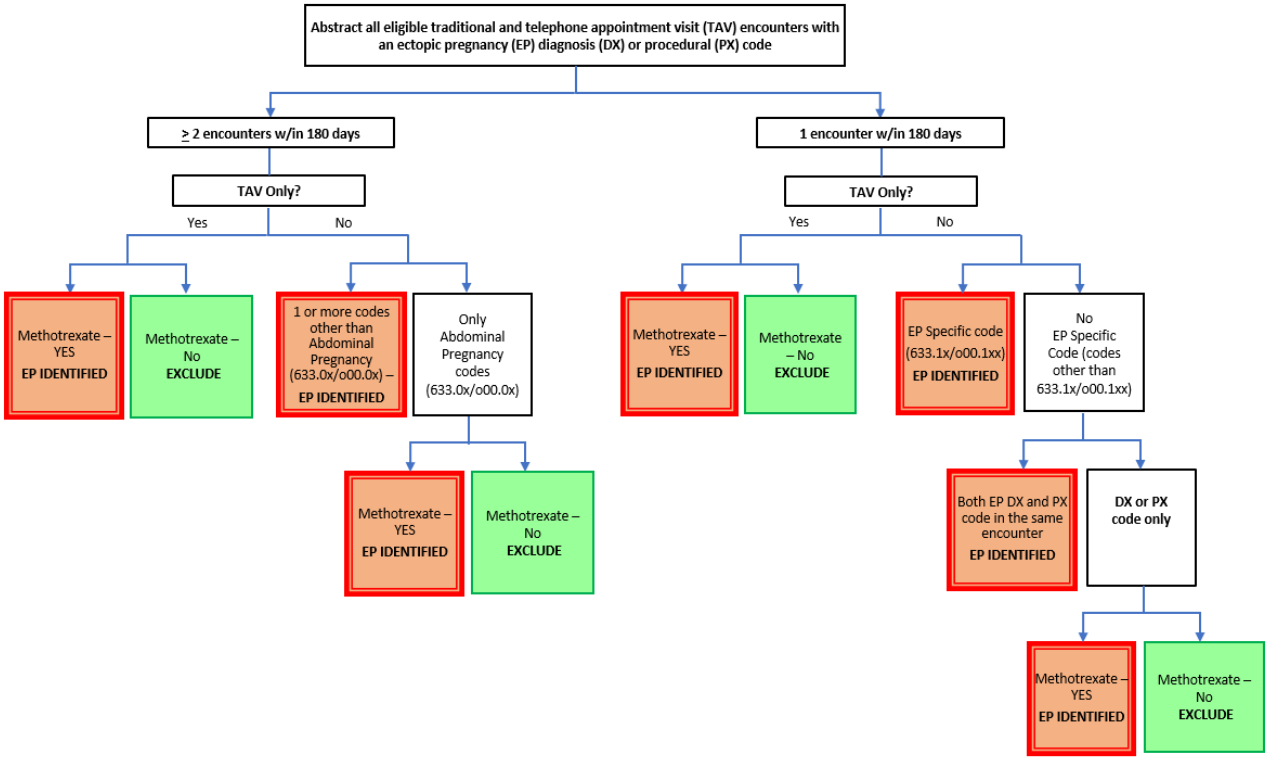
Kaiser Permanente Northern California (KPNC) and Southern California (KPSC) are the two largest Kaiser Permanente regions of the nine regional entities in the United States. These integrated health care systems provide health care service to over 9 million racially and ethnically diverse members who receive their care mainly from KP physicians and allied staff in 36 hospitals and over 427 medical centers scattered throughout California. Both KPSC and KPNC access the Virtual Data Warehouse, which was created to facilitate multi-site research projects. KP health care staff in both outpatient and inpatient clinical settings utilize an EHR based on an Epic platform that is accessible to multiple health care providers at the same time and in multiple locations. KPSC and KPNC fully implemented the EHR system for both outpatient care encounters and inpatient services in 2008 and 2009, respectively. It is a highly sophisticated integrated health information management and care management system designed to enhance the quality of patient care. The data is collected in real time with patient-centered records that provide access to comprehensive patient information to clinicians and researchers more instantly, efficiently, and securely compared with pre-EHR era paper records.

We developed an enhanced algorithm to identify EPs in the two health care systems through several iterative steps: First, we incorporated corresponding ICD-10 diagnostic and procedure codes that were not in use when the Scholes et al algorithm was developed in 2011. We then chart reviewed an initial random sample of 100 cases (50 KPNC and 50 KPSC) that had at least one EP diagnostic or procedure code but were not classified as EP by the Scholes et al algorithm to understand the reasons for misclassification. This information was used to modify the Scholes et al algorithm to improve the accuracy of case ascertainment. In addition to the inclusion of ICD-10 diagnostic/procedure codes, the major changes that were made to the previously validated algorithm as a result of our initial chart review were the addition of a new source of information (telephone appointment visits [TAVs]), the exclusion of cases with only abdominal EP diagnosis codes, additional criteria of a combination of an EP diagnostic and procedure code to be considered a case, refinement of methotrexate medication codes that were considered valid, and expansion of the allowable days from the assigned EP diagnosis date to administration of methotrexate.

The final enhanced algorithm (Figure 1) that was developed required either (1) at least 2 encounters, including at least 1

in-person visit, with an EP code other than abdominal EP (abdominal codes O00.00 and O00.01); (2) at least 2 TAVs with an EP code and evidence of methotrexate use; (3) at least 1 outpatient or inpatient visit or outside claims visit with any of the specific ICD, Ninth Revision (ICD-9), or ICD-10 diagnostic codes 633.10, 633.11, O00.10, and O00.11; (4) a combination of any single encounter (outpatient or inpatient visit, outside claims visit, or TAV) with a nonspecific EP code plus evidence of methotrexate use; or (5) a single non-TAV encounter with both an EP diagnosis and procedure code on the same encounter.

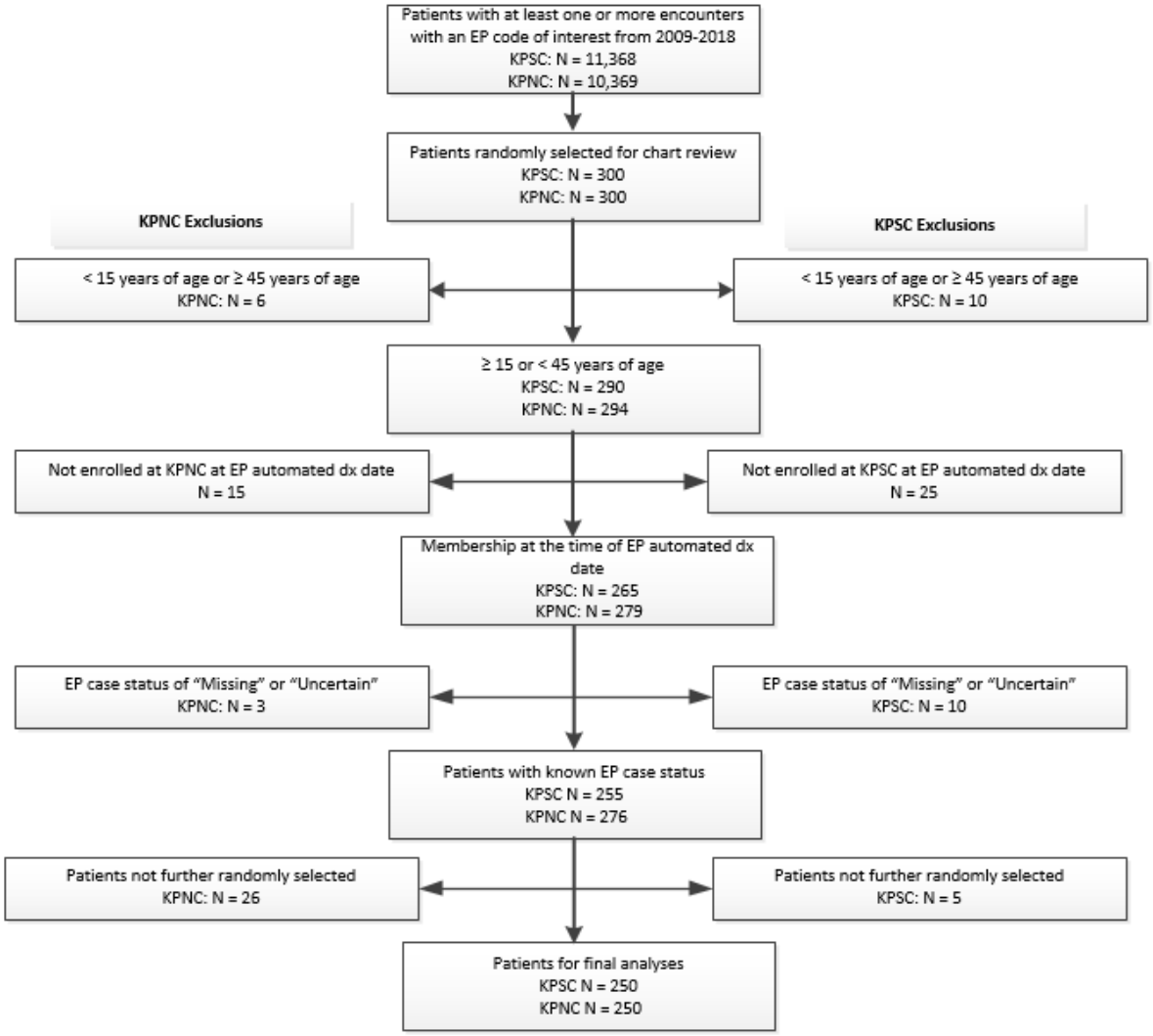
Figure 1. Enhanced ectopic pregnancy algorithm.



The EP diagnosis date was defined as the date of the first encounter with an EP code. Multiple encounters with EP codes occurring within a 180-day period from the first encounter with an EP code were considered part of the same pregnancy episode. Methotrexate use was defined as a medication code found within 30 days prior to and 180 days after the first EP diagnosis date. The justification for relaxing the criteria for methotrexate administration to 30 days prior to the first diagnosis, in contrast to the 7 days allowed in Scholes et al algorithm, was to minimize misclassification of treatment status due to inaccurate assignment of EP diagnosis dates. In randomly selected chart abstractions, we also found that methotrexate medication codes had various administrative subcodes that corresponded with true use of methotrexate; hence, we had to specify medication administration subcodes.

To assess the validity of the previously validated algorithm by Scholes et al and the newly developed enhanced version of the algorithm against the gold-standard "true case" as determined by chart review, a random sample of 600 patients (300 at each site) with a potential EP was selected. A potential case was defined as any case with at least 1 ICD-9, ICD-10, or Current Procedural Terminology code for EP (Multimedia Appendix 1). This approach was chosen because, in our setting, as in most health care settings that rely on insurance reimbursement, it is unlikely for an EP case to not have documentation with either a diagnosis or procedural code. Therefore, we assumed that cases that did not meet the initial inclusion criteria would be very unlikely to be a true EP case. By limiting the sample to cases with these inclusion criteria, we increased the number of true cases with little risk of missing cases. Further inclusion criteria were applied (women who were aged 15 to 44 years from January 1, 2009, to December 31, 2018, and were enrolled in the health plan for at least 1 month over the study period) to the 600 randomly selected cases. Cases that did not meet these requirements were excluded, leaving 255 cases at KPSC and 276 at KPNC. We randomly selected 250 cases from each site for chart review for this validation study (Figure 2).

Figure 2. Flow diagram of validation study sample. EP: ectopic pregnancy. KPNC: Kaiser Permanente Northern California. KPSC: Kaiser Permanente Southern California.



Using a standardized abstraction form, chart reviews were performed by trained abstractors to identify true EP cases. Cases where EP status was unclear were identified and adjudicated by a clinician. In our analysis of preliminary data pulls, we found that 10.5% (1568/14,907) of EP cases identified using the Scholes et al algorithm for classification could not be clearly classified as either medical or surgical. Therefore, information on treatment modality (surgical vs medical) was collected to assess the level of agreement. EP cases were classified as surgically managed if the patient had undergone any EP removal surgery within 30 days of the first encounter with an EP code, regardless of whether the patient received methotrexate. Remaining EP cases were classified as medically treated if the patient received methotrexate for an EP. Cases for which the type of treatment could not be determined were considered unclassified.

The test performance of both algorithms was calculated on the 500 potential EP cases: sensitivity (percentage of chart review–confirmed cases that were correctly classified as EP by the algorithm), specificity (percentage of cases determined not to be EP by chart review that were correctly classified by the algorithm), positive predictive value (PPV; percentage of cases classified as EP by the algorithm that were confirmed by chart review), and negative predictive value (NPV; percentage of identified cases classified as not EP by the algorithm that were determined not to be EP cases from chart review). Furthermore, the overall test performance of a dichotomous diagnostic test was assessed using the Youden J statistic [24] (Youden index=sensitivity+specificity–1), and the weighted harmonic mean of the test's precision and recall were assessed by computing the F-score (2×[PPV×sensitivity]/[PPV+sensitivity]). Agreement in case identification between the Scholes et al and the enhanced algorithms was assessed using kappa (κ) statistics. In addition, we evaluated the performance of electronic abstraction in correctly identifying EP management type (medical or surgical) among confirmed EP cases compared to that of chart review using the same performance measures. Lastly, we conducted a sensitivity analysis calculating the same performance measures using the Scholes et al algorithm and enhanced algorithm for a subset of cases from 2009 to the end of 2014 (ICD-9–only cases).

Results

Table 1 shows the distribution of maternal characteristics among the study sample and the two study sites (KPSC and KPNC) from which the sample for this validation study was drawn. Only a small proportion of the women in the sample population were teens and over a third were Hispanic. There was a higher proportion of Hispanic members at KPSC than at KPNC and a higher proportion of non-Hispanic White and Asian/Pacific Islander members at KPNC than at KPSC. Only a small proportion of women in the sampled cohort lived in neighborhoods with a median annual household income below US $30,000. Although the distribution of maternal characteristics is largely comparable between the sampled population and the overall cohort, women in the sampled population were slightly more likely to be from non-Hispanic Black backgrounds and less likely to be from non-Hispanic White racial/ethnic backgrounds.

Table 1. Characteristics of the validation study sample and the combined and site-specific populations.

Characteristics	Sample (n=500)	KPSC[a] and KPNC[b] populations		
	Chart reviewed, n (%)	Overall, n (%) (N=19,615)	KPSC, n (%) (n=9823)	KPNC, n (%) (n=9792)
Maternal age (years)				
<20	22 (4.4)	668 (3.4)	353 (3.6)	315 (3.2)
20-29	169 (33.8)	7036 (35.9)	3643 (37.1)	3393 (34.7)
30-34	157 (31.4)	6073 (31.0)	2970 (30.2)	3103 (31.7)
≥35	152 (30.4)	5838 (29.8)	2857 (29.1)	2981 (30.4)
Race/ethnicity				
Non-Hispanic White	124 (24.8)	5458 (27.8)	2257 (23.0)	3201 (32.7)
Non-Hispanic Black	82 (16.4)	2579 (13.1)	1298 (13.2)	1281 (13.1)
Hispanic	199 (39.8)	7668 (39.1)	4960 (50.5)	2708 (27.7)
Asian/Pacific Islander	83 (16.6)	3261 (16.6)	1069 (10.9)	2192 (22.4)
Other	4 (0.8)	349 (1.8)	144 (1.5)	205 (2.1)
Unknown	8 (1.6)	300 (1.5)	95 (1.0)	205 (2.1)
Smoking status[c]				
No	461 (92.2)	17,947 (91.5)	8929 (90.9)	9018 (92.1)
Yes	39 (7.8)	1668 (8.5)	894 (9.1)	774 (7.9)
Parity				
Nullipara	146 (29.2)	5690 (29.0)	2671 (27.2)	3019 (30.8)
Multipara	259 (51.8)	10,444 (53.2)	5214 (53.1)	5230 (53.4)
Missing/unavailable	95 (19.0)	3481 (17.7)	1938 (19.7)	1543 (15.8)
Family household income[d] (US $)				
<$30,000	31 (6.2)	1092 (5.6)	584 (5.9)	508 (5.2)
$30,000-$49,999	117 (23.4)	4863 (24.8)	2806 (28.6)	2057 (21.0)
$50,000-$69,999	147 (29.4)	5474 (27.9)	2913 (29.7)	2561 (26.2)
$70,000-$89,999	104 (20.8)	4131 (21.1)	1969 (20.0)	2162 (22.1)
≥$90,000	101 (20.2)	4033 (20.6)	1535 (15.6)	2498 (25.5)

[a]KPSC: Kaiser Permanente Southern California.

[b]KPNC: Kaiser Permanente Northern California.

[c]Smoking status documented within the year prior to the index date.

[d]Median family household income based on census tract of residence.

Chart review demonstrated that 334 (66.8%) of the 500 cases were true ectopic pregnancies. The sensitivity, specificity, PPV, and NPV of using the Scholes et al algorithm and the enhanced algorithm for identifying EPs are presented in Table 2. The sensitivity, specificity, NPV, and PPV for the Scholes et al algorithm were lower at 94.3% (315/334), 84.3% (140/166), 88.1% (140/159), and 92.4% (315/341), respectively, compared to those for the enhanced algorithm at 97.6% (326/334), 84.9%

(141/166), 94.6% (141/149), and 92.9% (326/351), respectively. Furthermore, the overall performance (Youden index and F-score) of the enhanced algorithm was higher than the performance of the Scholes et al algorithm at 82.5 and 95.2 versus 78.7 and 93.3, respectively.

Table 2. Ectopic pregnancy ascertainment performance of the Scholes et al and enhanced ectopic pregnancy algorithms.

Characteristic	Scholes et al algorithm			Enhanced algorithm		
	Yes	No	Total	Yes	No	Total
Classification by chart review, n						
Yes	315	19	334	326	8	334
No	26	140	166	25	141	166
Total	341	159	500	351	149	500
Test characteristics						
Sensitivity, % (n/N)	N/A[a]	N/A	94.3 (315/334)	N/A	N/A	97.6 (326/334)
Specificity, % (n/N)	N/A	N/A	84.3 (140/166)	N/A	N/A	84.9 (141/166)
Negative predictive value, % (n/N)	N/A	N/A	88.1 (140/159)	N/A	N/A	94.6 (141/149)
Positive predictive value, % (n/N)	N/A	N/A	92.4 (315/341)	N/A	N/A	92.9 (326/351)
Youden index	N/A	N/A	78.6	N/A	N/A	82.5
F-score	N/A	N/A	93.3	N/A	N/A	95.2

[a]N/A: not applicable.

We evaluated the performance of electronic abstraction in correctly identifying EP management type in the 326 EP cases identified by both the chart review and the enhanced algorithm. Chart review revealed that 197 (60.4%) were managed surgically, 126 (38.7%) were managed medically, and 3 (0.9%) could not be classified. Electronic abstraction assigned 186 (57.1%) EP cases as managed surgically and 124 (38.0%) as managed medically, and 16 (4.9%) could not be classified. The performance of electronic chart abstraction in assigning EP management compared to that of chart review is provided in Table 3. The sensitivity of surgical procedure codes from electronic chart abstraction to correctly identify surgical management was 91.9% (181/197). The overall accuracy, defined as the percentage of EP cases with correct management (surgical, medical, and unclassified) identified by electronic chart abstraction, was 92.3% (301/326). An excellent level of agreement in EP case identification (κ=0.93, 95% CI 0.89-0.96) was observed between the Scholes et al algorithm and the enhanced algorithm.

Table 3. Ectopic pregnancy management ascertainment performance of electronic data abstraction.

Characteristic	Classification by electronic abstraction[a]			
	Surgical	Medical	Unclassified	Total
Classification by chart review, n				
Surgical	181	5	11	197
Medical	5	118	3	126
Unclassified	0	1	2	3
Total	186	124	16	326[a]
Test characteristics				
Sensitivity, % (n/N)	91.9 (181/197)	N/A[b]	N/A	N/A
Specificity, % (n/N)	96.1 (124/129)	N/A	N/A	N/A
Negative predictive value, % (n/N)	88.6 (124/140)	N/A	N/A	N/A
Positive predictive value, % (n/N)	97.3 (181/186)	N/A	N/A	N/A
Youden index	88	N/A	N/A	N/A
F-score	94.5	N/A	N/A	N/A
Overall accuracy[c], % (n/N)	N/A	N/A	N/A	92.3 (301/326)

[a]Includes cases confirmed as ectopic pregnancy by chart review and the enhanced algorithm.

[b]N/A: not applicable.

[c]The percentage of ectopic pregnancy cases with correct management (surgical, medical, and unclassified) identified by electronic chart abstraction.

Sensitivity analysis limiting data to the subset of cases (n=307) from 2009 to 2014 with ICD-9–only codes revealed that the sensitivity and NPV for the Scholes et al subset analysis, at 94.5% (206/218) and 85.9% (73/85), respectively (Table 4), were similar to 94.3% (315/334) and 88.1% (140/159), respectively, for the Scholes et al full data set (Table 2). The performance of the enhanced algorithm in the subset analyses (sensitivity of 97.2%, 212/218; NPV of 92.4%, 73/79) was also similar to the performance of the enhanced algorithm for the full data set (sensitivity of 97.6%, 326/334; NPV of 94.6%, 141/149).

Table 4. Sensitivity analysis of ectopic pregnancy ascertainment performance of the Scholes et al [4] and enhanced ectopic pregnancy algorithms on a 2009-2014 ICD-9–only subset.

Characteristic	Scholes et al algorithm			Enhanced algorithm		
	Yes	No	Total	Yes	No	Total
Classification by chart review, n						
Yes	206	12	218	212	6	218
No	16	73	89	16	73	89
Total	222	85	307	228	79	307
Test characteristics						
Sensitivity, % (n/N)	N/A[a]	N/A	94.5 (206/218)	N/A	N/A	97.2 (212/218)
Specificity, % (n/N)	N/A	N/A	82.0 (73/89)	N/A	N/A	82.0 (73/89)
Negative predictive value, % (n/N)	N/A	N/A	85.9 (73/85)	N/A	N/A	92.4 (73/79)
Positive predictive value, % (n/N)	N/A	N/A	92.8 (206/222)	N/A	N/A	93.0 (212/228)
Youden index	N/A	N/A	76.5	N/A	N/A	79.3
F-score	N/A	N/A	93.6	N/A	N/A	95.1

[a]N/A: not applicable.

Discussion

In this validation study of EP, we found that our enhanced version of an algorithm that was previously validated by Scholes et al [4] in 2011 for identification of EP had a slightly higher sensitivity of 97.6% and negative predictive value of 94.6% compared to the original algorithm. The overall test performance, as estimated by the Youden index and F-score, was also much higher for the enhanced algorithm. However, we found similar specificities and PPVs in both the enhanced and Scholes et al algorithms. Furthermore, limiting the test performance to the pre-EHR era, the period when ICD-9 was used to code and classify medical conditions (2009-2014), the enhanced algorithm yielded a higher sensitivity, NPV, and overall test performance in EP case identification, suggesting that differences are due to improvement in clinical information collection and retrieval rather than any ICD code changes (from ICD-9 to ICD-10).

The quality of data extracted from outpatient encounters and hospital discharge records has been well studied. The accuracy of data abstraction varies by health care system, coding and clinical practice, and design of EHR query modules, among others [25]. For example, in a fee-for-service setting, in-person visits may be the primary mode of care; however, in a capitated care model, telephone encounters, which are not billable but allow providers to speak directly with patients who may be at home or another convenient location, may be used more frequently. These appointments usually last about 20 minutes and do not require a copay. Although an efficient option that helps patients avoid unnecessary in-person doctor visits, the usefulness and quality of data extracted from TAVs has not been well studied. We evaluated the performance of our enhanced algorithm after including TAV in the algorithm and found that accuracy improved when TAV EP codes were used in combination with EP codes from in-person encounters (Multimedia Appendix 2).

Scholes et al developed the original algorithm using a classification and regression tree (CART) [26]. The CART model is a nonparametric classification technique for building decision trees in which results are presented in a useful and easy-to-interpret “tree” format. However, it does not generate prediction probabilities needed to assess calibration. Model discriminatory accuracy is typically assessed. We made minor modifications to the algorithm to incorporate equivalent ICD-10 diagnostic and procedure codes and took into account other coding differences unique to the current EHRs (ie, new medication codes) and clinical practice (ie, increasing use of TAVs). Therefore, our enhanced algorithm is updated to a more current health care setting, has high PPV for case identification, and will support contemporary observational studies with validated accuracy. Since our enhanced algorithm had a higher PPV than the Scholes et al algorithm and the agreement with the Scholes algorithm was high, we did not perform a new CART analysis.

Accurate case identification using the enhanced algorithm is feasible and increasingly useful for public health disease surveillance and epidemiological studies. Furthermore, early identification of high-risk women may provide better opportunities for early detection of EP in affected women.

The overall accuracy of electronic data abstraction to identify surgical management of EP was 92.3%. Although we demonstrated a high overall accuracy using surgical codes (Table 3), consideration should also be given to using additional surgical codes for tubal surgery that were not included in the

case-finding algorithm because they were not EP-related codes but may be used by some providers at the time of EP surgery in order to increase the accuracy of management assignment.

This study has strengths and limitations. The socioeconomically diverse patient population at KPNC and KPSC, which is broadly representative of California, makes our findings widely generalizable to health systems with similar clinical patterns (ie, closed health care systems). However, future research is needed to examine whether the enhanced algorithm can be applied in other settings. The validation of the enhanced algorithm based on EHRs during the time periods both prior to and subsequent to EHR implementation further enhances the strength of this study. While we attempted to identify all potential EP cases by using cases with either an EP-related diagnostic or procedure code, it is possible that EP cases that were incorrectly or not coded were not captured, which would have falsely increased the sensitivity of both algorithms. We did not adjust for the influence of baseline characteristics. Therefore, some caution in interpreting the findings is warranted.

The enhanced algorithm yielded better overall EP case identification test results from EHR data, with slight improvements in sensitivity, specificity, and predictive values compared to the algorithm developed using pre-EHR era data, suggesting that the accuracy of EP case identification can be improved by supplementing the Scholes et al algorithm with TAV and ICD-10 diagnosis and procedure codes from EHRs. Overall, the enhanced algorithm for EP case identification in integrated health care databases is adequate to allow for its use in future epidemiological studies. Further studies on the quality of EHRs geared toward specific prenatal outcomes are urgently needed.

Acknowledgments

This study was funded by Bayer AG. We appreciate the contributions of the Kaiser Permanente members who provided their electronic health record information to this study.

Conflicts of Interest

DG is the Principal Investigator at the KPSC site. DG reports grants from Bayer AG during the conduct of the study and grants from Centers for Disease Control and Prevention and the US National Institutes of Health/National Institute of Child Health and Human Development (NIH/NICHD) outside the submitted work. MJF is a coinvestigator at the KPSC site. MJF reports grants from Bayer AG during the conduct of the study and grants from NIH/NICHD outside the submitted work. TRB is the Principal Investigator at the KPNC site and employed by KPNC. She reports grants from Bayer AG during the conduct of the study and outside the submitted work. MAA is a coinvestigator at the KPNC site and employed by KPNC. She also reports grants from Bayer AG during the conduct of the study and outside the submitted work. AA is an employee of Bayer AG, the sponsoring company of this study. AA reports stocks from Bayer AG. SA and MC are employed by KPNC and report grants from Bayer AG during the conduct of the study. JMS, VYC, FX, TMI, JS, and HST are employed by KPSC and report grants from Bayer AG during the conduct of the study. The opinions expressed are solely the responsibility of the authors and do not necessarily reflect the official views of the funding agency.

Multimedia Appendix 1
International Classification of Diseases diagnostic and procedure codes (ICD-9 and ICD-10), Current Procedural Terminology (CPT-4) codes for ectopic pregnancy in the enhanced algorithm*.
[DOCX File , 13 KB - medinform_v8i11e18559_app1.docx]

Multimedia Appendix 2
Ectopic pregnancy ascertainment in telephone appointment visits - performance of electronic data abstraction.
[DOCX File , 13 KB - medinform_v8i11e18559_app2.docx]

References

1. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. J Epidemiol Community Health 2014 Mar;68(3):283-287. [doi: 10.1136/jech-2013-202744] [Medline: 24248997]
2. Loffer FD. Outpatient management of ectopic pregnancies. Am J Obstet Gynecol 1987 Jun;156(6):1467-1472. [doi: 10.1016/0002-9378(87)90018-4] [Medline: 2954464]
3. Ory SJ. New options for diagnosis and treatment of ectopic pregnancy. JAMA 1992;267(4):534-537. [Medline: 1530874]
4. Scholes D, Yu O, Raebel MA, Trabert B, Holt VL. Improving automated case finding for ectopic pregnancy using a classification algorithm. Hum Reprod 2011 Nov;26(11):3163-3168 [FREE Full text] [doi: 10.1093/humrep/der299] [Medline: 21911435]
5. Steib SA, Reichley RM, McMullin ST, Marrs KA, Bailey TC, Dunagan WC, et al. Supporting ad-hoc queries in an integrated clinical database. Proc Annu Symp Comput Appl Med Care 1995:62-66 [FREE Full text] [Medline: 8563360]
6. Johnson SB, Hripcsak G, Chen J, Clayton P. Accessing the Columbia Clinical Repository. Proc Annu Symp Comput Appl Med Care 1994:281-285 [FREE Full text] [Medline: 7949935]

7. Byar DP. Problems with using observational databases to compare treatments. Stat Med 1991 Apr;10(4):663-666. [doi: 10.1002/sim.4780100417] [Medline: 2057663]
8. McDonald CJ, Hui SL. The analysis of humongous databases: problems and promises. Stat Med 1991 Apr;10(4):511-518. [doi: 10.1002/sim.4780100404] [Medline: 2057652]
9. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
10. Kemper AR, Uren RL, Clark SJ. Adoption of electronic health records in primary care pediatric practices. Pediatrics 2006 Jul;118(1):e20-e24. [doi: 10.1542/peds.2005-3000] [Medline: 16818534]
11. D'Avolio LW. Electronic medical records at a crossroads: impetus for change or missed opportunity? JAMA 2009 Sep 09;302(10):1109-1111. [doi: 10.1001/jama.2009.1319] [Medline: 19738097]
12. Dunn MJ. Benefits of electronic medical records outweigh every challenge. WMJ 2007 May;106(3):159-160 [FREE Full text] [Medline: 17642356]
13. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med 2009 Sep 01;151(5):359-360. [doi: 10.7326/0003-4819-151-5-200909010-00141] [Medline: 19638404]
14. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
15. Gallego B, Dunn AG, Coiera E. Role of electronic health records in comparative effectiveness research. J Comp Eff Res 2013 Nov;2(6):529-532. [doi: 10.2217/cer.13.65] [Medline: 24236790]
16. Anthony MS, Armstrong MA, Getahun D, Scholes D, Gatz J, Schulze-Rath R, et al. Identification and validation of uterine perforation, intrauterine device expulsion, and breastfeeding in four health care systems with electronic health records. Clin Epidemiol 2019;11:635-643 [FREE Full text] [doi: 10.2147/CLEP.S201044] [Medline: 31413641]
17. Smith N, Iyer RL, Langer-Gould A, Getahun DT, Strickland D, Jacobsen SJ, et al. Health plan administrative records versus birth certificate records: quality of race and ethnicity information in children. BMC Health Serv Res 2010 Nov 23;10:316 [FREE Full text] [doi: 10.1186/1472-6963-10-316] [Medline: 21092309]
18. Andrade SE, Scott PE, Davis RL, Li D, Getahun D, Cheetham TC, et al. Validity of health plan and birth certificate data for pregnancy research. Pharmacoepidemiol Drug Saf 2013 Jan;22(1):7-15 [FREE Full text] [doi: 10.1002/pds.3319] [Medline: 22753079]
19. Coleman KJ, Ngor E, Reynolds K, Quinn VP, Koebnick C, Young DR, et al. Initial validation of an exercise "vital sign" in electronic medical records. Med Sci Sports Exerc 2012 Nov;44(11):2071-2076. [doi: 10.1249/MSS.0b013e3182630ec1] [Medline: 22688832]
20. Paul DW, Neely NB, Clement M, Riley I, Al-Hegelan M, Phelan M, et al. Development and validation of an electronic medical record (EMR)-based computed phenotype of HIV-1 infection. J Am Med Inform Assoc 2018 Feb 01;25(2):150-157 [FREE Full text] [doi: 10.1093/jamia/ocx061] [Medline: 28645207]
21. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013 Jun;20(e1):e147-e154 [FREE Full text] [doi: 10.1136/amiajnl-2012-000896] [Medline: 23531748]
22. Smith N, Coleman KJ, Lawrence JM, Quinn VP, Getahun D, Reynolds K, et al. Body weight and height data in electronic medical records of children. Int J Pediatr Obes 2010 May 03;5(3):237-242. [doi: 10.3109/17477160903268308] [Medline: 19961272]
23. Getahun D, Rhoads G, Fassett M, Chen W, Strauss J, Demissie K, et al. Accuracy of reporting maternal and infant perinatal service system coding and clinical utilization coding. J Med Stat Inform 2013;1:1-3 [FREE Full text] [doi: 10.7243/2053-7662-1-3]
24. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]
25. Horsky J, Drucker EA, Ramelson HZ. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. AMIA Annu Symp Proc 2017;2017:912-920 [FREE Full text] [Medline: 29854158]
26. Breiman L, Friedman JH, Stone CJ, Olshen RA. Classification and Regression Trees. Belmont, CA: Taylor and Francis; 1984.

Abbreviations

CART: classification and regression tree
EHRs: electronic health records
EP: ectopic pregnancy
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
KPNC: Kaiser Permanente Northern California

KPSC: Kaiser Permanente Southern California
NIH/NICHD: National Institutes of Health/National Institute of Child Health and Human Development
NPV: negative predictive value
PPV: positive predictive value
TAVs: telephone appointment visits

Edited by G Eysenbach; submitted 04.03.20; peer-reviewed by M Bannick, R Bajpai; comments to author 12.06.20; revised version received 23.07.20; accepted 30.10.20; published 30.11.20.

Please cite as:
Getahun D, Shi JM, Chandra M, Fassett MJ, Alexeeff S, Im TM, Chiu VY, Armstrong MA, Xie F, Stern J, Takhar HS, Asiimwe A, Raine-Bennett T
Identifying Ectopic Pregnancy in a Large Integrated Health Care Delivery System: Algorithm Validation
JMIR Med Inform 2020;8(11):e18559
URL: http://medinform.jmir.org/2020/11/e18559/
doi:10.2196/18559
PMID:33141678

Original Paper

Exploring Fever of Unknown Origin Intelligent Diagnosis Based on Clinical Data: Model Development and Validation

Huizhen Jiang[1*], MSc; Yuanjie Li[2*], MD; Xuejun Zeng[2], MD; Na Xu[2], MD; Congpu Zhao[1], MSc; Jing Zhang[1], BSc; Weiguo Zhu[1,2], MD

[1]Department of Information Center, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Department of Primary Care and Family Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Weiguo Zhu, MD
Department of Primary Care and Family Medicine
Peking Union Medical College Hospital
Chinese Academy of Medical Sciences and Peking Union Medical College
1 Shuaifuyuan
Dongcheng District
Beijing, 100730
China
Phone: 86 010 69154149
Email: Zhuwg@pumch.cn

Abstract

Background: Fever of unknown origin (FUO) is a group of diseases with heterogeneous complex causes that are misdiagnosed or have delayed diagnoses. Previous studies have focused mainly on the statistical analysis and research of the cases. The treatments are very different for the different categories of FUO. Therefore, how to intelligently diagnose FUO into one category is worth studying.

Objective: We aimed to fuse all of the medical data together to automatically predict the categories of the causes of FUO among patients using a machine learning method, which could help doctors diagnose FUO more accurately.

Methods: In this paper, we innovatively and manually built the FUO intelligent diagnosis (FID) model to help clinicians predict the category of the cause and improve the manual diagnostic precision. First, we classified FUO cases into four categories (infections, immune diseases, tumors, and others) according to the large numbers of different causes and treatment methods. Then, we cleaned the basic information data and clinical laboratory results and structured the electronic medical record (EMR) data using the bidirectional encoder representations from transformers (BERT) model. Next, we extracted the features based on the structured sample data and trained the FID model using LightGBM.

Results: Experiments were based on data from 2299 desensitized cases from Peking Union Medical College Hospital. From the extensive experiments, the precision of the FID model was 81.68% for top 1 classification diagnosis and 96.17% for top 2 classification diagnosis, which were superior to the precision of the comparative method.

Conclusions: The FID model showed excellent performance in FUO diagnosis and thus would be a potentially useful tool for clinicians to enhance the precision of FUO diagnosis and reduce the rate of misdiagnosis.

(JMIR Med Inform 2020;8(11):e24375) doi:10.2196/24375

KEYWORDS

fever of unknown origin; intelligent diagnosis; machine learning; BERT; fever; misdiagnosis

Introduction

Fever is one of the most common symptoms in medicine [1]. A febrile temperature may boost the immune system to fight disease [2]. Prolonged fevers are usually complex to diagnose [3]. A fever of unknown origin (FUO) has remained a challenging diagnostic problem in recent decades [4].

As there are more than 200 causes of FUO [3], isolating the cause of an FUO is a great challenge for clinicians. Thus, many clinicians have been drawn to FUO research [5]. In 1961, Petersdorf and Beeson [6] defined FUO. There are usually 3 characteristics: (1) prolonged fever for more than 3 weeks, (2) recurrent fever with a temperature higher than 38.3℃, and (3) undiagnosed fever after a 1-week inpatient investigation [6,7]. The definition has been revised over time with regard to the classification and the duration of fever to be diagnosed [8,9]. The classification of FUO has been hotly debated in previous studies [10,11]. Usually, the categories of causes are infections, immune diseases, and tumors [12,13], and their treatment methods are considerably different, including anti-infection medication, hormones, and chemotherapy, respectively. Therefore, if the cause of an FUO is diagnosed to one category, regardless of the disease causing the FUO, the treatment direction can basically be determined, which would be meaningful for doctors. Infection is the most common cause of FUO [14]. Chow and Robinson [15] found that for more than one-half of children with FUO, the FUO was caused by an infection. Knockaert et al [9] proposed that the wait-and-see strategy may be better for prolonged prognosis in adults. Therefore, FUO is worth studying and exploring. Previous studies on FUO have mainly analyzed real patients' cases. de Kleijn and von der Meer [16] assessed 53 patients with FUO using a statistical analysis tool called PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [17]. de Kleijn et al [18] analyzed 167 patients with FUO using fixed criteria. Similarly, Efstathiou et al [13] discriminated FUO into infectious and noninfectious causes. While these research results might explain how these patients with FUO were diagnosed, it was difficult to automatically determine the method because the limited amount of data sometimes caused overfitting. Little research on FUO has been done using machine learning methods.

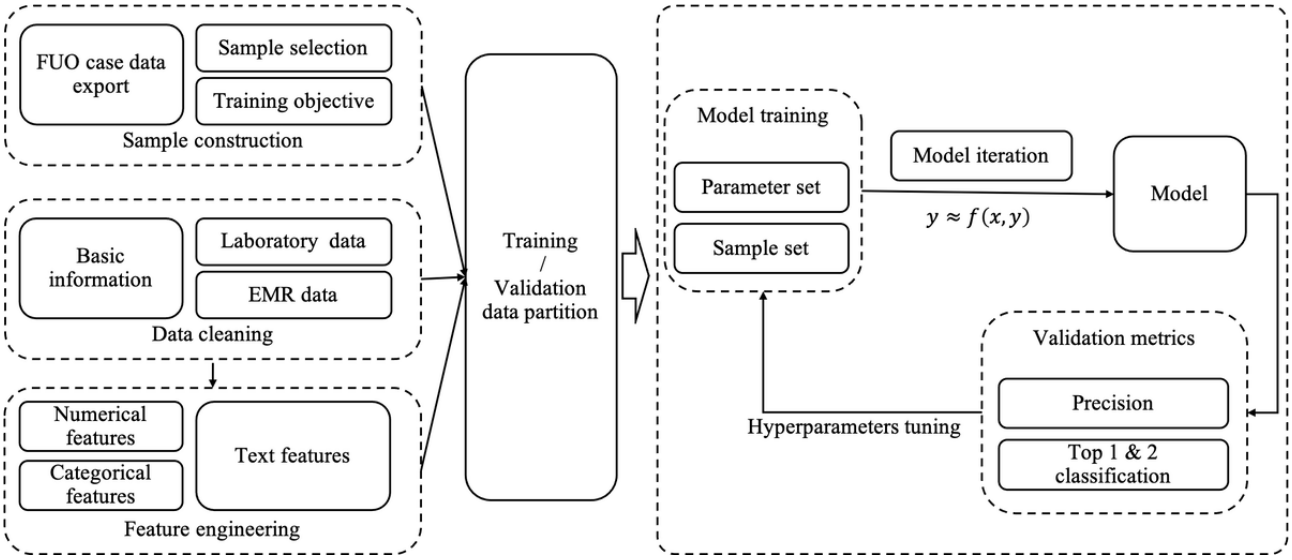
In this paper, we proposed an FUO intelligent diagnosis (FID) model to classify the causes of FUO into 4 categories based on clinical data. Extensive experiments showed good performance of the model. In summary, we made the following contributions: (1) we innovatively introduced an FUO intelligent classification diagnosis model called FID, which can automatically group FUO cases into one of the categories of causes, (2) our experiments were based on real, desensitized data from Peking Union Medical College Hospital and thus, the cases were real and the results were more valuable and credible when applied to a clinical setting, and (3) we conducted extensive experiments to evaluate the performance of the FID based on the gradient boosting methods LightGBM and XGBoost, which performed better on small data sets; the FID model achieved better performance using LightGBM.

Methods

Modeling

In this paper, we proposed the FID model using LightGBM to intelligently diagnose patients with FUO into 1 of 4 causes using their basic information, clinical laboratory data, and electronic medical record (EMR) data. The structure of the FID model is shown in Figure 1. First, we classified the causes of FUO into 4 categories: infections, immune diseases, tumors, and others. Then, we cleaned the basic information and laboratory data and structured the EMR data via 2 methods: the bidirectional encoder representations from transformers (BERT) model [19] and Jieba [20]. Next, we extracted the features and trained the model through extensive experiments. Finally, by comparing the experimental results, we evaluated the performance of the FID model.

Figure 1. Structure of the fever of unknown origin (FUO) intelligent diagnosis (FID) model. The left side shows the sample construction, data cleaning, and feature engineering process; the right side shows the model training and tuning processes. EMR: electronic medical record.



Data Sources

This research was based on data from 2299 desensitized FUO cases from Peking Union Medical College Hospital from June 1, 2012, to March 31, 2018, and the data filtering process is shown in Figure 2. The data contained basic information, laboratory results, and EMR data. One patient visit was taken as 1 sample; if a patient was admitted to the hospital twice, then the 2 visits were taken as 2 samples. There were 3723 total cases whose chief complaint included "fever." First, we filtered out the 52 cases not eligible for FUO diagnosis, such as those with temperatures lower than 38.3°C. Then, we invited 3 doctors specializing in FUO diagnosis to help check the data and classifications. Two doctors divided the remaining cases into 4 categories: infections, immune diseases, tumors, and others. The third doctor checked the classifications, and if there were disagreements, the 3 doctors discussed the cases until they obtained a consistent classification. Based on the doctors' suggestion, 1372 cases that had no confirmed diagnosis were filtered out. As a result, 2299 cases whose causes fit into 1 of the 4 categories remained for the experiments. Infections, immune diseases, tumors, and others accounted for 52.28% (1202/2299), 36.50% (839/2299), 7.83% (180/2299), and 3.39% (78/2299) of cases, respectively. The distribution of the data is shown in Figure 3. We randomly divided 80% of the data into the training set and the remaining 20% into the validation set, and we used the validation set data to evaluate the performance of the model.

Figure 2. Data filtering process.

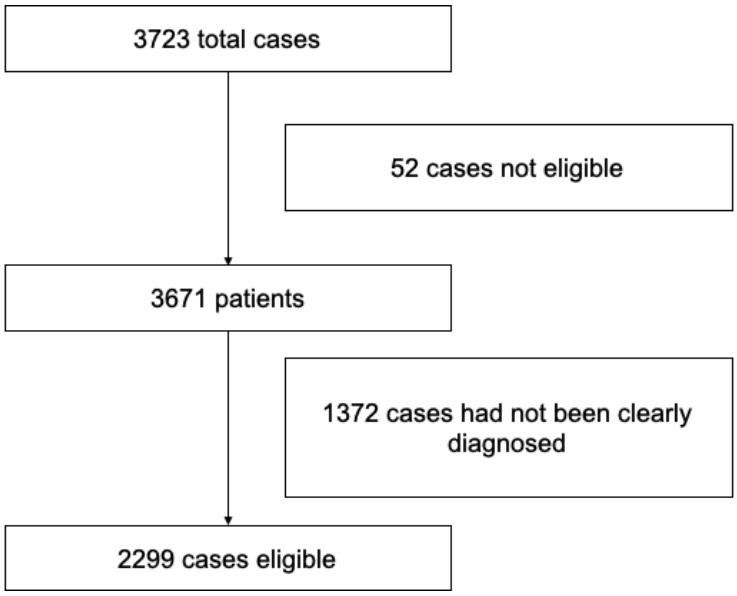
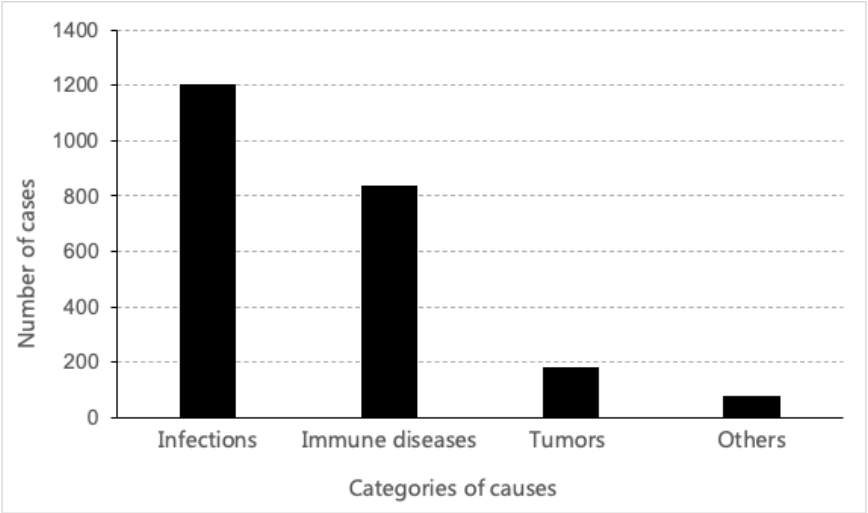


Figure 3. Distribution of the data set.



Model

Sample Structure

After obtaining the data marked by doctors, we needed to clean the basic information and laboratory data and obtain the structured EMR data. The EMR data were unstructured. To structure the EMR data, we used the BERT model. The BERT model was proposed by Devlin et al [19] in 2018, and it has greatly improved the text structuring process. We used the BERT-based pretrained model to process the text in this study. The input of BERT was all of the EMR text and the text in each line belonging to 1 patient. Each line in the output of BERT

was a number vector of 768 dimensions. In addition, we compared the results based on BERT and Jieba. Using Jieba, we segmented all the text data and chose the top 100 text segmentations according to the counts of the words that occurred in all the text, such as “fever,” “infected,” and “lymph node.” In addition, stop words were filtered out manually. Then, for each of the 100 text segmentations, if it existed in the EMR text of 1 patient, we used 1 to represent the segmentation; otherwise, we used 0. Finally, the text of each patient was expressed by a vector of 100 dimensions.

Data Cleaning

The data were irregular after being extracted from the clinical system. Regarding the laboratory results, each item group consisted of too many cell items; therefore, laboratory data could be a thousand dimensions. However, there were some items that occurred only a few times, which were difficult for the model to learn. Therefore, we filtered out the items occurring less than 10 times. Finally, 214 items were left for the laboratory data. Then, we transformed all the values and units to be consistent. Regarding the synonyms, we fixed them to be the same. For example, “1 L” and “1000 ml” were fixed to “1000 mL.”

Feature Engineering

Feature engineering is the most important part of machine learning. The performance of a model is determined by feature engineering to a large extent. In this paper, there were 3 kinds of features: numerical features, categorical features, and text features.

Numerical features were objective data, such as age and heart rate. For the numerical features, we directly extracted the values as the features.

Categorical features were the classification indicators, such as gender and positive and negative symptoms. Categorical features were addressed using the LabelEncoder method [21] with numbers starting from 0, and then numbers were assigned to the features.

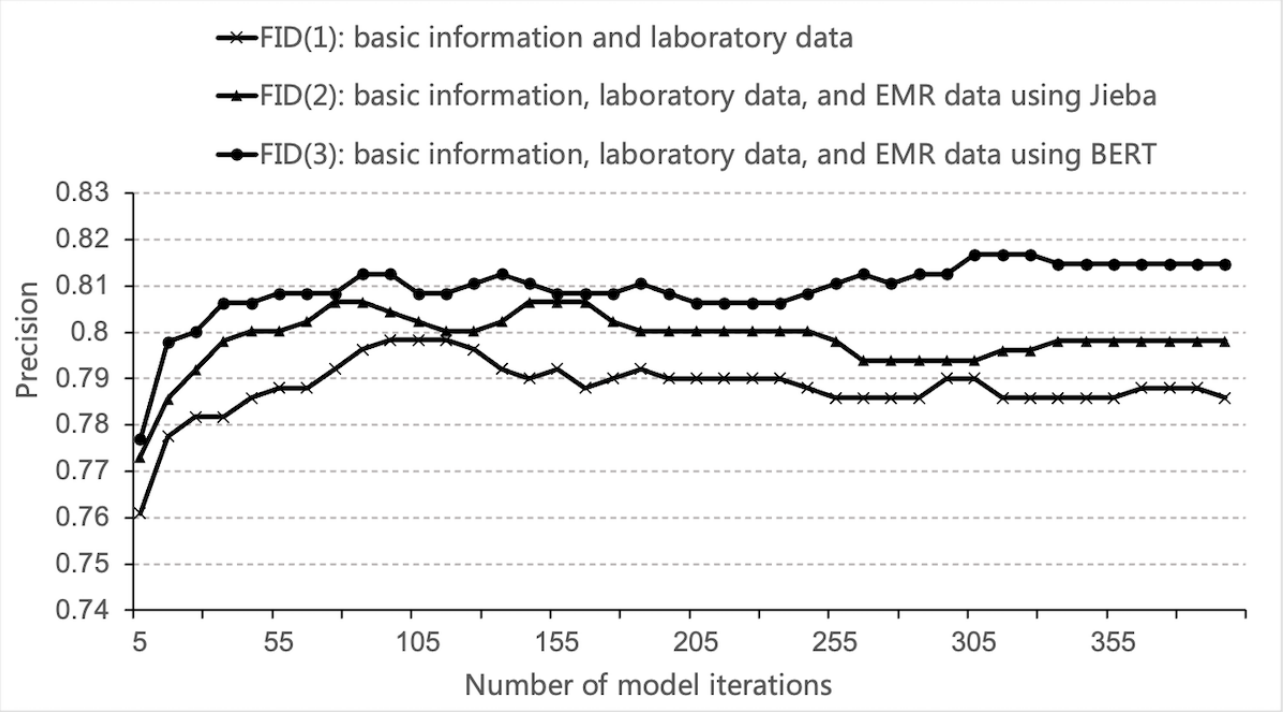
Text features, which could also contribute necessary information, were mainly from the EMR. They were structured in the previous step. Then, we merged the structured EMR features with all of the features. We used 2 methods to structure the text: using Jieba, the top 100 segmentations were selected as the features according to the counts of every word segment, and using BERT, all the text was structured as 768 features.

Results

We used the LightGBM framework to train the model. LightGBM is a gradient boosting framework that is based on decision trees and has faster training efficiency, lower memory usage, and higher precision than other frameworks, as well as a large-scale data processing capability. During the training of LightGBM, there were many parameters that needed to be optimized, including the number of iterations, the learning rate, and the number of leaf nodes. During the experiments, we found that different parameters impacted the final results of the model to some extent. In this research, the parameters we used were a learning rate of 0.01 and number of leaves of 6 after the experiments. The output of the multiclassification model was the probability that each sample belonged to each category. Therefore, the classification with the largest prediction probability was generally regarded as the sample classification.

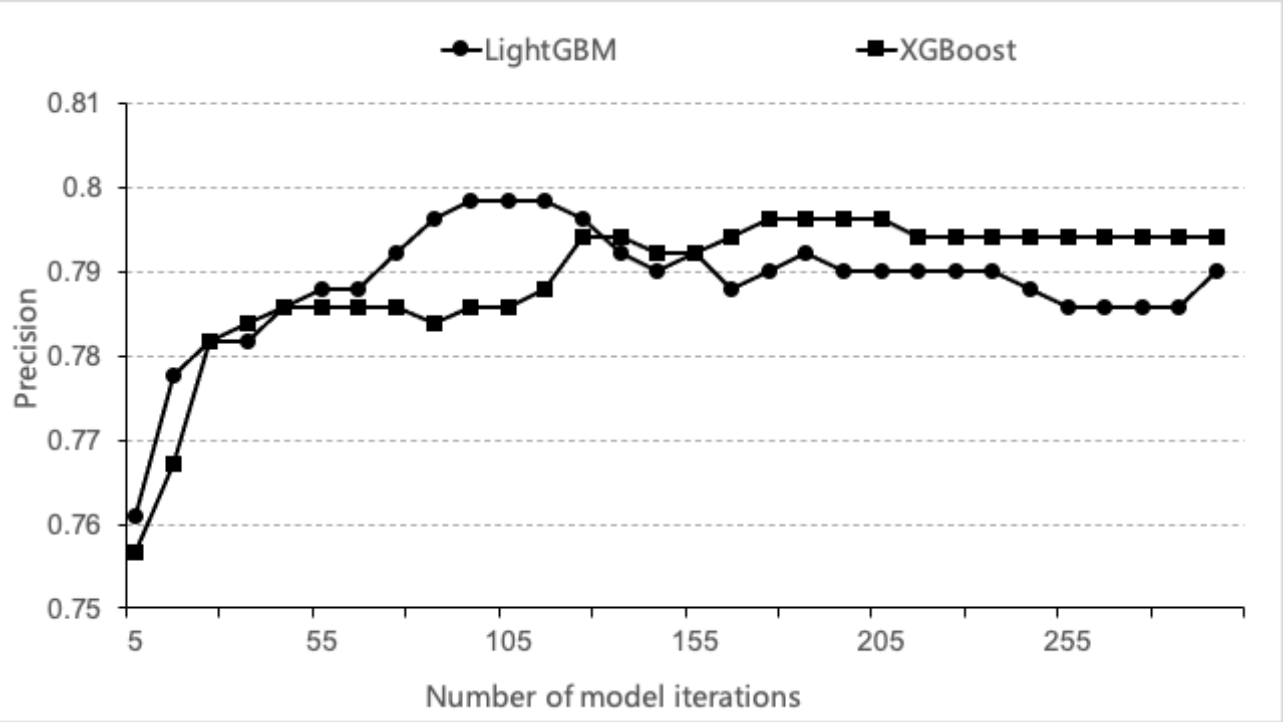
Based on different features, we built 3 models: FID(1) was based on the basic information and laboratory features; FID(2) was based on the basic information, laboratory data, and EMR data using Jieba; and FID(3) was based on the basic information, laboratory data, and EMR data using BERT. The results are shown in Figure 4. The figure shows that the precision was further improved by increasing the number of features from the text data. After introducing the text features developed by BERT, we obtained the optimal model: FID(3). The precision was 81.68%. Experiments show that, in addition to structured data, unstructured text data also hides a lot of valuable information, which could improve the performance of the model.

Figure 4. Performance of different data sets. FID: fever of unknown origin intelligent diagnosis. EMR: electronic medical record. BERT: bidirectional encoder representations from transformers.



In addition to the LightGBM algorithm, we also tested XGBoost, an algorithm with a similarly good performance as LightGBM. Both are gradient boosting algorithms. As shown in Figure 5, the precision of LightGBM was higher than that of XGBoost with relatively fewer iterations. When the number of iterations increased, the precision of XGBoost exceeded that of LightGBM, but it was still lower than the best performance of LightGBM. Therefore, we used LightGBM as the training algorithm for the subsequent experiments in this study.

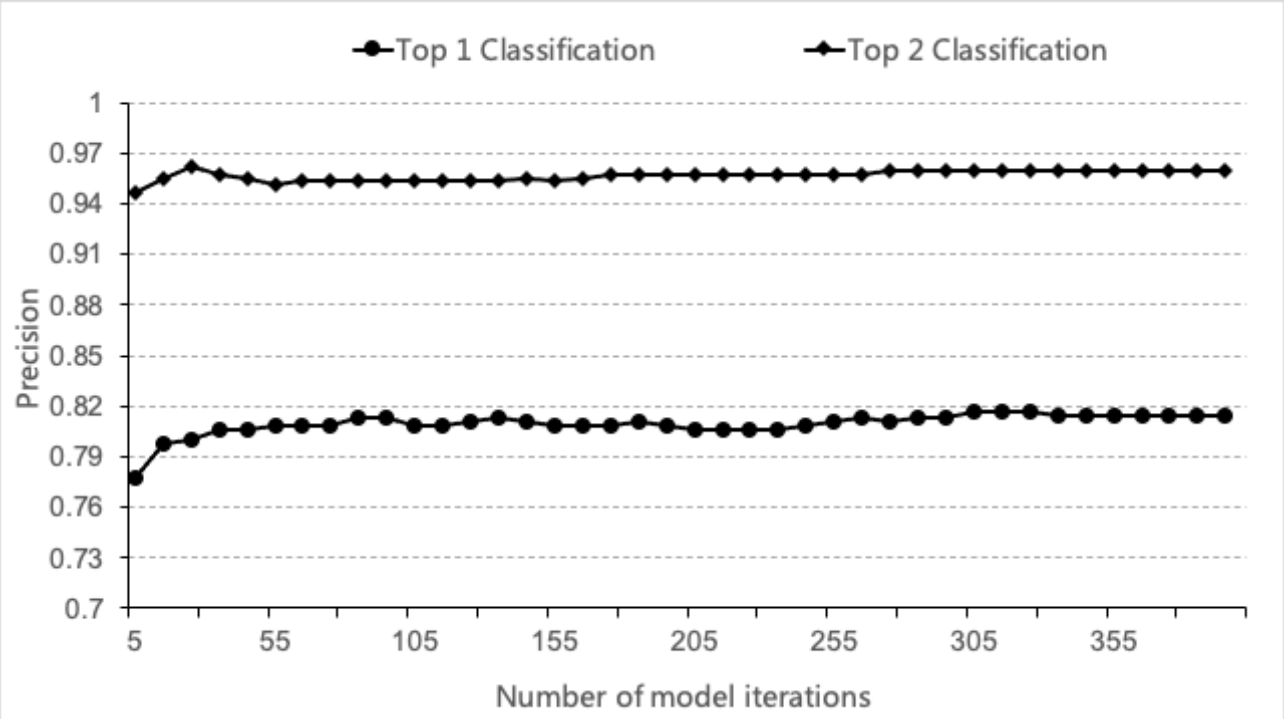
Figure 5. Performance based on the LightGBM and XGBoost algorithms. The abscissa represents the number of model iterations, and the ordinate shows the precision of the model.



In fact, in many cases, it is difficult for a model to make a completely accurate decision. The greatest value of a model is that it can provide more suggestions for decision making. If the classification with the largest prediction probability is not accurate but the top 2 classifications are accurate, it could also provide considerable help for doctors. Here, we evaluated the

precision of the 2 classifications with the largest prediction probabilities. In Figure 6, we can see that the precision of the first 2 categories of the model was 96.17% and there were few mistakes.

Figure 6. Performance of top 1 and top 2 classifications.



In addition, we explored the patient distribution according to average age and gender, as shown in Table 1. We observed that for FUO caused by tumors, the average age of patients was the highest at 50.91 years. Regarding FUO caused by immune diseases, the gender distribution showed a large difference, with 34.93% (803/2299) of patients being male and 65.07% (1496/2299) of patients being female. For the other 3 categories of FUO, there were no obvious differences in gender.

Table 1. Patient distribution in the data set.

	Patient demographics (N=2299)		
Categories of cause of fever of unknown origin	Mean age (years)	Male, n (%)	Female, n (%)
Infections	47.79	1226 (53.33)	1073 (46.67)
Immune diseases	42.90	803 (34.93)	1496 (65.07)
Tumors	50.91	1201 (52.24)	1098 (47.76)
Others	38.81	1268 (55.15)	1031 (44.85)

Usually, before doctors diagnose the cause of a patient's FUO, they need to make an appointment to examine the patient. Table 2 shows the top 10 laboratory measurements that were associated with an FUO cause diagnosis. Percentage of basophils was the measurement that showed the strongest correlation with the FUO cause diagnosis. The other 9 measurements were percentage of large unstained cells, age, fibrinogen level, thrombin time, alkaline phosphatase level, direct bilirubin level, blood sodium level, 24-hour urine volume, and lymphocyte count. The top 10 measurements could be provided to doctors as laboratory appointment decision support.

Table 2. Top 10 laboratory measurements related to the diagnosis of the cause of fever of unknown origin.

Number	Top 10 laboratory measurements
1	Percentage of basophils
2	Percentage of large unstained cells
3	Age
4	Fibrinogen level
5	Thrombin time
6	Alkaline phosphatase level
7	Direct bilirubin level
8	Blood sodium level
9	24-hour urine volume
10	Lymphocyte count

Discussion

Principal Findings

Intelligent diagnosis of the cause category of FUO is significant and practical. With the rapid development of information technology, big data has been the focus of many fields in recent years [22]. Volume, variety, velocity, and value are the 4 "V" characteristics, although mining the deep value of data sets is the most important aspect for big data research. Similarly, in the medical field, large amounts of health care data are produced every day, such as EMRs, laboratory results, and images [23]. Considerable amounts of precious information can be extracted and mined from medical data using the proper methods. Previously, experts manually identified and analyzed the meaning of health data [24], which was time-consuming and difficult to identify specifically. Medical big data is increasingly more accepted by doctors because of its high efficiency and lower costs [25]. Rodger [26] mentioned that not only was a data extraction system needed, but medical big data applications such as those for clinical decision support were in urgent demand. Medical big data applications help make the medical process easier and friendlier for patients and relieve the pressure on clinicians. Among patients with FUO, the proportion of undiagnosed patients is approximately 20.5% [13]. In particular, the treatment for FUO may be much different, even contrary, for different causes. Therefore, helping doctors discover the specific cause as soon as possible is meaningful.

We addressed the problem using more appropriate and superior methods. Currently, medical big data applications have explored many directions [27,28]. The different kinds of methods used can be divided into 4 types: data mining, image recognition, natural language processing (NLP), and speech recognition. For example, intelligent diagnosis with the data excluding images [29,30] and intelligent early warnings [31] are both data mining problems. For image recognition [32], Simonyan and Zisserman [32] examined very deep convolutional networks and achieved superior performance. NLP is mainly used in the structured analysis of EMRs [19,33]. In this research, NLP also played an important role in structuring the text data, and we processed the EMR data using BERT and Jieba. Speech recognition [34], mainly addressed using recurrent neural networks, could help doctors transfer voice to text with high efficiency. Regarding medical big data models, there are unsupervised learning models and supervised learning models [35], such as logistic regression, decision tree, deep learning, and others. Currently, supervised learning models are used more often because of the sensitivity of the data to medical knowledge. As most problems in medical big data are classification problems, decision trees and deep learning models could achieve better performance. Wu et al [36] exploited the diagnosis of hypocellular myelodysplastic syndrome and aplastic anemia, and their experiments showed that the decision tree model outperformed the others in classification. Most importantly, deep learning methods require very large data sets [37], usually millions of data sets. Since there were only 2299 cases in this study, gradient boosting methods were better for this research. Therefore, LightGBM and XGBoost were used to train the data. In addition, 1372 cases not clearly diagnosed were removed from the study, and this kind of case would exist in the real-world setting. Therefore, the precision might be lower in reality, and these cases should be taken into account in future work.

Conclusion

A machine learning method was innovatively introduced into FUO diagnosis. We presented the FUO intelligent diagnosis model called the FID, which was based on basic information, laboratory data, and EMR data from Peking Union Medical College Hospital. After cleaning the disordered data and structuring the text data using BERT, we conducted many experiments on the sample data and compared the performances from several angles. The results showed that the FID outperformed the comparative methods. As the treatments for FUO from different causes are very different, intelligently diagnosing an FUO into a category is meaningful. Our research was based on data from 1 hospital, and we intelligently diagnosed the FUO to 1 category of causes. In the future, we will focus on predicting the exact cause of an FUO using multicenter data. We would include all cases, including cases with no confirmed diagnosis, in our future research to better match real-world scenarios, which would probably improve the method more practical for the real clinical process.

Acknowledgments

HJ, YL, and WZ contributed to the study concept and design. CZ, YL, XZ, JZ, and NX contributed to the acquisition of the data set. HJ, YL, and WZ consulted on the analyses. All authors interpreted the results, contributed to the manuscript, and approved the final draft.

This study was supported by the National Key Research and Development Program of China (project 2018YFC0116905) and the CAMS Innovation Fund for Medical Sciences (CIFMS) project 2016-I2 M-2-004.

Conflicts of Interest

None declared.

References

1. Ifesinachi P. Mechanisms of fever in humans. International Journal of Microbiology and Immunology Research 2013:43 [FREE Full text]
2. Evans SS, Repasky EA, Fisher DT. Fever and the thermal regulation of immunity: the immune system feels the heat. Nat Rev Immunol 2015 May 15;15(6):335-349. [doi: 10.1038/nri3843]
3. Cunha BA, Lortholary O, Cunha CB. Fever of Unknown Origin: A Clinical Approach. The American Journal of Medicine 2015 Oct;128(10):1138.e1-1138.e15. [doi: 10.1016/j.amjmed.2015.06.001]
4. Keidar Z, Gurman-Balbir A, Gaitini D, Israel O. Fever of Unknown Origin: The Role of 18F-FDG PET/CT. Journal of Nuclear Medicine 2008 Nov 07;49(12):1980-1985. [doi: 10.2967/jnumed.108.054692]
5. Cunha CB. Prolonged and Perplexing Fevers in Antiquity: Malaria and Typhoid Fever. Infectious Disease Clinics of North America 2007 Dec;21(4):857-866. [doi: 10.1016/j.idc.2007.08.010]
6. Petersdorf RG, Beeson PB. Fever of unexplained origin: Report on 100 cases. Medicine 1961;40(1):1-30. [doi: 10.1097/00005792-196102000-00001] [Medline: 13734791]
7. Fusco FM, Pisapia R, Nardiello S, Cicala SD, Gaeta GB, Brancaccio G. Fever of unknown origin (FUO): which are the factors influencing the final diagnosis? A 2005-2015 systematic review. BMC Infect Dis 2019 Jul 22;19(1):653 [FREE Full text] [doi: 10.1186/s12879-019-4285-8] [Medline: 31331269]
8. Durack DT, Street AC. Fever of unknown origin--reexamined and redefined. Curr Clin Top Infect Dis 1991;11:35-51. [Medline: 1651090]
9. Knockaert DC, Vanderschueren S, Blockmans D. Fever of unknown origin in adults: 40 years on. J Intern Med 2003 Mar;253(3):263-275. [doi: 10.1046/j.1365-2796.2003.01120.x]
10. Pasic S, Minic A, Djuric P, Micic D, Kuzmanovic M, Sarjanovic L, et al. Fever of unknown origin in 185 paediatric patients: A single-centre experience. Acta Paediatrica 2006 Apr 1;95(4):463-466. [doi: 10.1080/08035250500437549]
11. Zhou W, Tan X, Li Y, Tan W. Human Herpes Viruses Are Associated with Classic Fever of Unknown Origin (FUO) in Beijing Patients. PLoS ONE 2014 Jul 3;9(7):e101619. [doi: 10.1371/journal.pone.0101619]
12. Wright WF, Auwaerter PG. Fever and Fever of Unknown Origin: Review, Recent Advances, and Lingering DogmaC. Open Forum Infectious Diseases. US: Oxford University Press 2020;7(5). [doi: 10.1093/ofid/ofaa132]
13. Efstathiou SP, Pefanis AV, Tsiakou AG, Skeva II, Tsioulos DI, Achimastos AD, et al. Fever of unknown origin: Discrimination between infectious and non-infectious causes. European Journal of Internal Medicine 2010 Apr;21(2):137-143. [doi: 10.1016/j.ejim.2009.11.006]
14. Burzo ML, Antonelli M, Pecorini G, Favuzzi AM, Landolfi R, Flex A. Fever of unknown origin and splenomegaly. Medicine 2017;96(50):e9197. [doi: 10.1097/md.0000000000009197]
15. Chow A, Robinson JL. Fever of unknown origin in children: a systematic review. World J Pediatr 2010 Dec 30;7(1):5-10. [doi: 10.1007/s12519-011-0240-5]
16. de Kleijn EM, van der Meer JW. Fever of unknown origin (FUO): report on 53 patients in a Dutch university hospital. Neth J Med 1995 Aug;47(2):54-60. [doi: 10.1016/0300-2977(95)00037-n] [Medline: 7566282]
17. Transparent reporting of systematic reviews and meta-analyses. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). URL: http://prisma-statement.org/PRISMAStatement/HistoryAndDevelopment [accessed 2020-11-24]
18. de Kleijn EMHA, Vandenbroucke JP, van der Meer JWM. Fever of Unknown Origin (FUO): I. A prospective multicenter study of 167 patients with FUO, using fixed epidemiologic entry criteria. Medicine 1997;76(6):392-400. [doi: 10.1097/00005792-199711000-00002]
19. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online May 24, 2019. [FREE Full text]

20. Peng KH, Liou LH, Chang CS, Lee D. Predicting personality traits of Chinese users based on Facebook wall posts. 2015 Presented at: Proceedings of the 2015 24th Wireless and Optical Communication Conference (WOCC); October 23-24, 2015; Taipei, Taiwan p. 9-14. [doi: 10.1109/WOCC.2015.7346106]
21. Bisong E. Introduction to Scikit-learn. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Berkeley, CA: Apress; 2019:215-229.
22. Chen M, Mao S, Liu Y. Big Data: A Survey. Mobile Netw Appl 2014 Jan 22;19(2):171-209. [doi: 10.1007/s11036-013-0489-0]
23. Lee CH, Yoon H. Medical big data: promise and challenges. Kidney Res Clin Pract 2017 Mar 31;36(1):3-11. [doi: 10.23876/j.krcp.2017.36.1.3]
24. Siuly S, Zhang Y. Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis. Data Sci. Eng 2016 Jul 27;1(2):54-64. [doi: 10.1007/s41019-016-0011-3]
25. Roski J, Bo-Linn GW, Andrews TA. Creating Value In Health Care Through Big Data: Opportunities And Policy Implications. Health Affairs 2014 Jul;33(7):1115-1122. [doi: 10.1377/hlthaff.2014.0147]
26. Rodger JA. Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive. Informatics in Medicine Unlocked 2015;1:17-26. [doi: 10.1016/j.imu.2016.01.002]
27. Sun J, Reddy CK. Big data analytics for healthcare. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013 Presented at: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining; 2013; Chicago, Illinois, USA p. 1525. [doi: 10.1145/2487575.2506178]
28. Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R. Big data knowledge system in healthcare. In: Bhatt C, Dey N, Ashour AS, editors. Internet of things and big data technologies for next generation healthcare. Heidelberg, Germany: Springer; 2017:157.
29. Jia F, Lei Y, Lin J, Zhou X, Lu N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mechanical Systems and Signal Processing 2016 May;72-73:303-315. [doi: 10.1016/j.ymssp.2015.10.025]
30. Avci E, Turkoglu I. An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases. Expert Systems with Applications 2009 Mar;36(2):2873-2878. [doi: 10.1016/j.eswa.2008.01.030]
31. Su H, Wen Z, Wu Z. Study on an Intelligent Inference Engine in Early-Warning System of Dam Health. Water Resour Manage 2011 Jan 15;25(6):1545-1563. [doi: 10.1007/s11269-010-9760-3]
32. Simonyan K, Zisserman A. Very deep convolutional networks for large- scale image recognition. arXiv Preprint posted online April 10, 2015. [FREE Full text]
33. Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. J Am Med Inform Assoc 2011 Sep 01;18(5):539-539. [doi: 10.1136/amiajnl-2011-000501]
34. Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. In: IEEE. 2013 Presented at: 2013 IEEE international conference on acoustics, speech and signal processing; 2013; Vancouver, BC, Canada. [doi: 10.1109/icassp.2013.6638947]
35. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Jun 21;2(4):230-243. [doi: 10.1136/svn-2017-000101]
36. Wu J, Zhang L, Yin S, Wang H, Wang G, Yuan J. Differential Diagnosis Model of Hypocellular Myelodysplastic Syndrome and Aplastic Anemia Based on the Medical Big Data Platform. Complexity 2018 Nov 12;2018:1-12. [doi: 10.1155/2018/4824350]
37. Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture 2018 Apr;147:70-90. [doi: 10.1016/j.compag.2018.02.016]

Abbreviations

BERT: bidirectional encoder representations from transformers
EMR: electronic medical record
FID: fever of unknown origin intelligent diagnosis
FUO: fever of unknown origin
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach; submitted 16.09.20; peer-reviewed by B Qian, L Wang; comments to author 06.10.20; revised version received 18.10.20; accepted 28.10.20; published 30.11.20.

Please cite as:
Jiang H, Li Y, Zeng X, Xu N, Zhao C, Zhang J, Zhu W
Exploring Fever of Unknown Origin Intelligent Diagnosis Based on Clinical Data: Model Development and Validation
JMIR Med Inform 2020;8(11):e24375
URL: http://medinform.jmir.org/2020/11/e24375/
doi:10.2196/24375
PMID:33172835

Original Paper

The Associations of Electronic Health Record Usability and User Age With Stress and Cognitive Failures Among Finnish Registered Nurses: Cross-Sectional Study

Anu-Marja Kaihlanen[1], PhD; Kia Gluschkoff[1], PhD; Hannele Hyppönen[1], PhD; Johanna Kaipio[2], PhD; Sampsa Puttonen[3], PhD; Tuulikki Vehko[1], PhD; Kaija Saranto[4], PhD; Liisa Karhe[5], PhD; Tarja Heponiemi[1], PhD

[1]Finnish Institute for Health and Welfare, Helsinki, Finland

[2]Department of Computer Science, Aalto University, Helsinki, Finland

[3]Finnish Institute of Occupational Health, Helsinki, Finland

[4]Department of Health and Social Management, University of Eastern Finland, Kuopio, Finland

[5]Finnish Nurses Association, Helsinki, Finland

Corresponding Author:
Anu-Marja Kaihlanen, PhD
Finnish Institute for Health and Welfare
P.O. Box 30
Helsinki, FI-00271
Finland
Phone: 358 295246033
Email: anu.kaihlanen@thl.fi

Abstract

Background: Electronic health records (EHRs) are expected to provide many clinical and organizational benefits. Simultaneously, the end users may face unintended consequences, such as stress and increased cognitive workload, due to poor EHR usability. However, whether the effects of usability depend on end user characteristics, such as career stage or age, remains poorly understood.

Objective: The objective of this study was to examine the associations of EHR usability and user age with stress related to information systems and cognitive failures among registered nurses.

Methods: A cross-sectional survey design was employed in Finland in 2017. A total of 3383 registered nurses responded to the nationwide electronic survey. Multiple linear regression was used to examine the associations of EHR usability (eg, how easily information can be found and a patient's care can be documented) and user age with stress related to information systems and cognitive failures. Interaction effects of EHR usability and age were also tested. Models were adjusted for gender and employment sector.

Results: Poor EHR usability was associated with higher levels of stress related to information systems (β=.38; *P*<.001). The strength of the association did not depend on user age. Poor EHR usability was also associated with higher levels of cognitive failures (β=.28; *P*<.001). There was a significant interaction effect between age and EHR usability for cognitive failures (β=.04; *P*<.001). Young nurses who found the EHR difficult to use reported the most cognitive failures.

Conclusions: Information system stress due to poor EHR usability afflicts younger and older nurses alike. However, younger nurses starting their careers may be more cognitively burdened if they find EHR systems difficult to use compared to older nurses. Adequate support in using the EHRs may be particularly important to young registered nurses, who have a lot to learn and adopt in their early years of practice.

(JMIR Med Inform 2020;8(11):e23623) doi:10.2196/23623

KEYWORDS

electronic health records; usability; stress; cognitive failure; nurse

Introduction

Electronic health record (EHR) systems have increasingly replaced paper-based practices in hospitals, with the expectation of providing many clinical and organizational benefits [1,2]. Implementation of EHRs will inevitably change work practices in health care [3,4] and, if not properly managed, may result in many unexpected and unintended consequences. In addition to the benefits identified (eg, reduction of medication errors), health care professionals have reported disadvantages, such as increased emotional strain and increased errors when trying to learn and adapt to new technologies and managing their workflow disruptions [2,5,6]. Implementation and use of EHRs have also required increased effort from professionals in performing their typical task flow [7], which in turn has resulted in increased cognitive workload and decreased cognitive performance [8]. In addition to missing focus on proactive workflow redesign during EHR implementation, earlier studies have found that many of the unfavorable consequences are connected with the usability issues of EHR systems, referring to how easy the system is to use and how precisely and efficiently required tasks can be performed [9,10].

Nurses, as the largest group of health care professionals, are the main end users of the EHRs, and their daily work is greatly influenced by ease of use as well as technical and functional quality of the systems. According to previous studies, both Finnish nurses and physicians are dissatisfied with the usability of their EHR systems [11]. Nurses' experiences of the poor usability of EHRs and how they can negatively affect workflow appear to be consistent across countries [12,13]. Ease of use and high quality of information systems promote the use of technology and support the management of care records [14]. However, EHRs that are perceived as difficult to use have been shown to be associated with increased stress levels [15] and cognitive workload among nurses [16,17], which may consequently increase the risk of cognitive failures [18]. Cognitive failure is defined as "a cognitively based error that occurs during the performance of a task that a person is normally successful in executing" [19]. The failure can occur in a person's memory functions, attention regulation, or actions [20], and the incidence is connected with work environment–related [18,21] and individual factors [20].

So far, only limited and contradictory knowledge exists on whether the consequences of poor EHR usability, particularly stress or impairment in cognitive functions, could depend on end user characteristics such as age. First, age-group differences in the likelihood of experiencing EHR-related stress have not been found [22]. However, experienced (and thus likely older) nurses can have more negative attitudes toward the use of new technologies in clinical work [23]. They may potentially experience more stress and higher cognitive workload than younger professionals, who may adapt better to digitization-related changes [16]. Second, youth has been associated with higher nursing informatics competence, such as skills in electronic documentation and use of information technology [24], which can make working with the EHRs easier and be a factor in protecting nurses' well-being at work [15]. Nevertheless, while young nurses starting their careers may be more skilled in and used to using technology than older nurses, they are still probably less experienced in using EHR systems in their work. Multiple studies have suggested that integrating EHRs into daily workflow and performing EHR-related tasks may be easier and require less cognitive effort from more experienced EHR users than from novice users [17,25].

Based on our knowledge, there is little evidence of whether nurses of a certain age are more at risk of experiencing stress or increased cognitive workload due to poor usability of EHR systems. In light of previous evidence, differences may exist between nurses of different ages, but the findings are mixed and the potential moderating effect of EHR usability has not been investigated. Moreover, previous studies examining the negative outcomes linked to EHR usability, such as stress, have mainly focused on physicians [26-28] and less on nurses [22]. Identifying those most at risk of experiencing EHR-related disadvantages is important in order to provide them with adequate support in using the systems. Most importantly, the topic requires further investigation because problems related to EHR usability and subsequent issues of stress and impairment in cognitive performance are notable threats to the quality of care and patient safety [18,29,30]. This study aimed to investigate the associations of EHR usability and user's age with stress related to information systems and cognitive failure at work among registered nurses. Additionally, we examined whether the possible associations of EHR usability with stress related to information systems and cognitive failures are modified by user age.

Methods

Setting, Data Collection, and Participants

In 2017, a nationwide cross-sectional survey was conducted in Finland on registered nurses' experiences with currently used EHR systems [15,31,32]. The data were collected with a web-based questionnaire that was sent to all the registered nurses (n=29,283) who were members of the Finnish Nurses Association and the National Association of Health and Welfare Professionals and had provided an email address. Altogether, 3607 of the 29,283 nurses responded to the questionnaire (a 12.3% response rate). Nurses with missing information on any of the demographic variables (age, gender, employment sector) were excluded from the study (n=224), resulting in a final sample of 3383 nurses. In the Finnish public health care system, the EHR coverage has been 100% since 2010 [33]. Over 20 EHR brands are being used in different health and social care settings [31].

Measurements

The EHR usability was measured with 7 ease of use–related items (α=.84) from the validated National Usability-Focused Health Information System Scale (NuHISS) [34]. The items were as follows: (1) the arrangement of fields and functions is logical on a computer screen; (2) the systems keep me clearly informed about what it is doing (eg, saving data); (3) terminology on the screen is clear and understandable (eg, titles and labels); (4) routine tasks can be performed in a straightforward manner without the need for extra steps using the system; (5) it is easy to obtain necessary patient information

using the information system; (6) entering and documenting patient data is quick, easy, and smooth; and (7) the information on the nursing record is in an easily readable format. Items were rated on a 5-point scale (1=fully disagree to 5=fully agree). A higher score on ease of use items indicates better experienced usability.

Stress related to information systems was measured with 2 items (α=.62) that evaluated how often during the past half-year period the person has been distracted, worried, or stressed about (1) constantly changing information systems and (2) difficult, poorly performing information technology equipment or software [26]. The items were rated on a 5-point scale ranging from 1 (never) to 5 (very often). This measure has been previously used with physicians and is associated with EHR usability and distress [27,35].

Cognitive failures were measured with 3 items (α=.59) modified from the 15-item Workplace Cognitive Failure Scale (WCFS) [20,36]. The WCFS includes 3 dimensions: failure in memory, failure in attention, and failure in action. Regarding the length of the survey questionnaire, we had to limit the number of questions and chose 1 item per dimension to measure nurses' cognitive failures. The selection of these items was based on their highest loadings for the 3 factors or dimensions of cognitive failure [20]. Participants were asked to rate how often they have faced situations at work where they (1) have not remembered a work-related password, set of numbers, etc (memory failure); (2) have not fully listened to the instructions or requests they have received (attention failure); or (3) have accidentally started or closed the wrong device, system, or program (action failure). Items were answered on a 5-point scale ranging from 1 (never) to 5 (several times a day).

Other variables included were age, gender, and employment sector (1=hospital, 2=health center, 3=private sector, 4=social services, or 5=other).

Data Analysis

Continuous variables are summarized using mean and standard deviation, and categorical variables are presented as the number of participants and percentage. Multiple linear regression was used to examine the associations of EHR usability and nurse's age with stress related to information systems and cognitive failures. This method was chosen because the associations were assumed to be linear, and it offered easily interpreted output coefficients and a less complex algorithm compared to many other methods. Analyses were conducted separately for both dependent variables (stress related to information systems and cognitive failures). In the first step, EHR usability and age were included as predictors in the model. In the second step, the combined effect of EHR usability and age was tested by adding an interaction term to the former model. Age was divided by 10 for the analysis to assess a given decade's association and make the estimated coefficients easier to interpret. All models were adjusted for gender and employment sector. The analyses were conducted using RStudio.

Results

Characteristics of the Participants

The majority of the participants were female (3204/3383, 94.7%). They were, on average, 46.2 years old (range: 22-66), and over half of the participants worked in hospitals. There were differences in the estimated EHR usability (P<.001), stress related to information systems (P<.001), and cognitive failures (P=.02) between nurses working in different work environments. The EHR usability was rated highest among nurses who worked in social services. Nurses working in hospitals gave the lowest EHR usability ratings and had more stress related to information systems than nurses working in other fields. There were no gender differences in the values of the variables studied. Characteristics of the participants and descriptive statistics of the study variables are presented in Table 1.

Table 1. Characteristics of the participants (N=3383) and descriptive statistics.

Characteristic	n (%)	Mean	SD	Minimum	Median	Maximum
Age, years	N/A[a]	46.22	11.1	22	48	66
Gender						
Male	169 (5.0)	N/A	N/A	N/A	N/A	N/A
Female	3204 (94.7)	N/A	N/A	N/A	N/A	N/A
Other	10 (0.3)	N/A	N/A	N/A	N/A	N/A
Employment sector						
Hospital	1796 (53.1)	N/A	N/A	N/A	N/A	N/A
Health center	707 (20.9)	N/A	N/A	N/A	N/A	N/A
Private clinic	173 (5.1)	N/A	N/A	N/A	N/A	N/A
Social services	433 (12.8)	N/A	N/A	N/A	N/A	N/A
Other	274 (8.1)	N/A	N/A	N/A	N/A	N/A
Usability	N/A	3.04	0.78	1	3	5
SRIS[b]	N/A	3.02	0.89	1	3	5
Cognitive failures	N/A	1.96	0.51	1	2	5

[a]N/A: not applicable.

[b]SRIS: stress related to information systems.

Associations of EHR Usability and User Age With Stress Related to Information Systems and Cognitive Failures

The results of the linear regression analyses are shown in Table 2. The EHR usability was associated with both stress related to information systems (β=.38; *P*<.001) and cognitive failures (β=.28; *P*<.001). Higher levels of usability were associated with lower levels of both stress related to information systems and cognitive failures. Age was associated with cognitive failures (β=.16; *P*<.001) but not with stress related to information systems. Younger nurses had higher levels of cognitive failure compared to older nurses.

There was a significant interaction effect between age and EHR usability for the cognitive failures (β=.04; *P*<.001). Younger nurses who evaluated the EHR as difficult to use had the highest levels of cognitive failures. Among older nurses, usability was not associated with their cognitive failure levels (Figure 1). There was no interaction effect between age and EHR usability for the stress related to information systems (Figure 2). Figure 1 and Figure 2 illustrate the interaction effects.

Table 2. The associations of age and EHR usability with stress related to information systems and cognitive failures.

Variable	Estimate	*P* value
SRIS[a]		
Age	.10	.15
Gender	.13	.10
Employment sector		
Hospital	Reference	N/A[b]
Health center	.10	.03
Private clinic	.32	<.001
Social service	.18	<.001
Other	.16	.03
Usability	.38	<.001
Age × usability	.01	.80
Adjusted R^2	0.16	N/A
Cognitive failures		
Age	.16	<.001
Gender	.02	.74
Employment sector		
Hospital	Reference	N/A
Health center	.08	.01
Private clinic	.03	.60
Social service	.00	.93
Other	.04	.33
Usability	.28	<.001
Age × usability	.04	<.001
Adjusted R^2	0.04	N/A

[a]SRIS: stress related to information systems.

[b]N/A: not applicable.

Figure 1. Interaction effect between EHR usability (ease of use) and user's age for cognitive failures. The association is shown for low (mean − 1 SD), average, and high (mean + 1 SD) levels of ease of use. EHR: electronic health record.

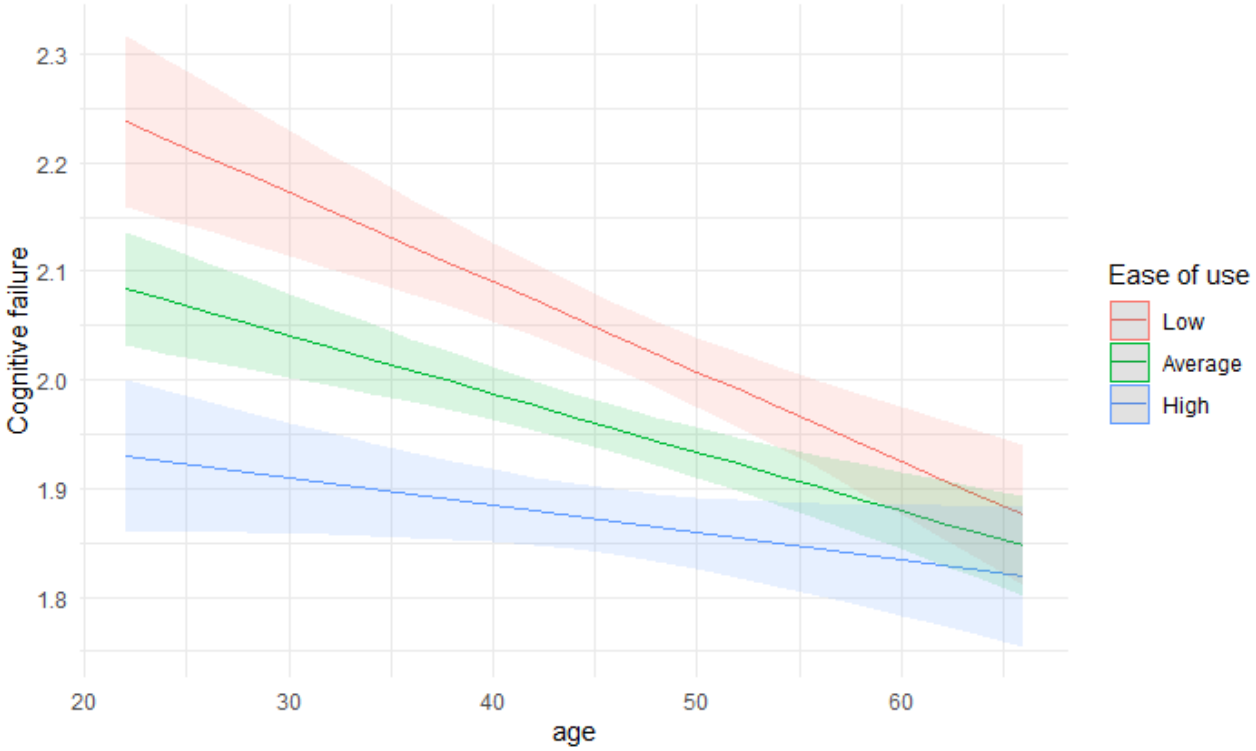
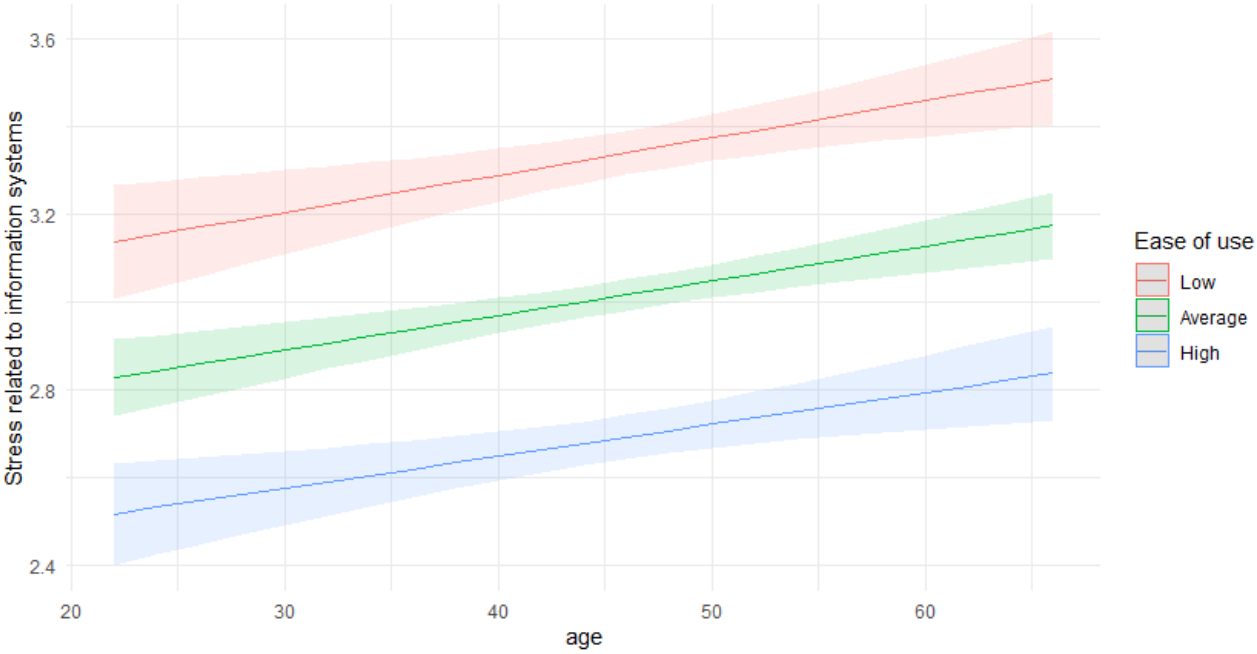


Figure 2. Interaction effect between EHR usability (ease of use) and user's age for information system–related stress. The association is shown for low (mean − 1 SD), average, and high (mean + 1 SD) levels of ease of use. EHR: electronic health record.



Discussion

Principal Findings

This study examined the associations of EHR usability and user's age with stress related to information systems and cognitive failures among Finnish registered nurses. The practical goal was to increase system vendors', health care managers', and nursing educators' awareness of the potential consequences of poor EHR usability, a shared problem among nurses in many countries that may jeopardize the quality and safety of care [12,29]. As predicted, we found that EHR usability was

associated with both stress related to information systems and cognitive failures. Nurses who provided higher usability ratings had less stress and fewer cognitive failures compared to nurses providing lower ratings. The nurses' age, in turn, was not associated with stress related to information systems, but it was negatively associated with cognitive failures. We also found a significant interaction effect between age and EHR usability for cognitive failures, indicating that young nurses who rated EHRs as difficult to use had the most cognitive failures.

The finding that young nurses' cognitive functions may be more impaired by poor EHR system usability compared to those of older nurses seems logical. The youngest nurses have probably worked in the field for the shortest time, and the early stages of a nursing career are known to be demanding and challenging, which can lead to their attentiveness and memory being strained [37,38]. In addition, older nurses with more experience of using EHRs and various EHR brands may find it easier to integrate information systems into daily workflow than less experienced nurses [25]. A previous study has shown conflicting results and found a relationship between higher age and increased risk of cognitive failures among nurses [21]. However, this result was explained by the fact that older nurses may have lower work ability, which is associated with an increased risk of cognitive failures in certain work environments [21]. Differences in work environments also emerged in this study, and it appears that EHRs may be particularly burdensome for nurses working in hospital settings, who found the systems most difficult to use and experienced the most stress compared to nurses working in other environments. It would be important to find out the views of nurses working in the hospital environment about the weaknesses of EHRs and how systems should be developed to improve their perceived usability and thereby reduce stress.

Due to new role adjustment, duties, responsibilities, and work environments, many new nurses experience increased stress and emotional exhaustion [39]. These symptoms have been associated with deterioration in cognitive performance (eg, in attention and memory functions), and this applies especially to professions with high levels of work pressure and intense cognitive demands, like nursing [40]. Although we only looked at stress related to information systems in this study, it is possible that the high level of strain early in their careers may partly explain why young nurses in this study had more cognitive failures. Moreover, due to lack of expertise and the fact that the youngest generation of nurses are the most likely to change jobs [41,42], they constantly have much to learn at work. This, on top of a load caused by EHRs that are difficult to use, may increase task stressors, such as performance constraints, task uncertainty, or difficulties managing time pressure and frequent interruptions, all of which are shown to foster cognitive failure [43].

It is evident that EHRs should support nurses in carrying out their work tasks and not in turn increase workload, stress, or cognitive burden. A recent review by Wisner et al concluded that EHRs have the potential to support the cognitive work of health care staff, but the scattering of information, information complexity, and lack of chronology often hampers this. Encountering problems while trying to find or synthesize information can affect a nurse's ability to achieve and maintain clinical understanding and situational awareness, which can compromise patient safety [8]. Usability and stability of information systems as well as end user involvement in system development and work procedure planning may be significant factors in alleviating stress related to information systems [15,26]. Since improving the usability of EHR systems seems to be challenging, the importance of adequate orientation and support at work to use information systems is critical.

In this study, challenges were observed especially in young nurses, whereby there is also a need to discuss whether current nursing education provides students with adequate knowledge and skills on how to use and integrate EHRs into daily work. Shortcomings have been identified in both theoretical and practical studies and, for example, in students' opportunities to practice documentation with real EHR systems during their education [44,45]. The fact that using the systems can often only be learned and practiced in the workplace after graduation puts an additional burden on young nurses when there is still a lack of mastery of the work and nursing as a whole. Moreover, the large number of different EHR brands and their differences in usage logic, for example, may slow the process of learning to use them.

Currently, work tasks that require the use of EHRs, such as documentation, take up a significant portion of nurses' day-to-day working hours [46]. Potential time pressure can be alleviated by having a high-quality information system [15]. Our study suggests that young nurses in particular could benefit from well-designed and implemented EHR systems that support routine tasks (eg, easy access to the information needed to treat the patient). Another interesting finding was that while poor usability of EHRs was associated with nurses' higher stress related to information systems, the level of stress did not vary significantly between younger and older nurses. In other words, the levels of tolerance of EHR usability problems appeared to be equal in nurses of different ages. The results of this study contradict the stereotypical idea that millennial nurses who have grown up with digitization and who are more accustomed to coping with a variety of electronic platforms and tasks simultaneously [47] would automatically be less burdened by information systems than those nurses who have had to learn to work with them at a later age. Older nurses may compensate for the slower adoption of information systems with their experience and better management of patients' overall care.

Limitations

Possible limitations of this study should be considered when interpreting the results. First, in spite of representativeness of the responses in a large sample of Finnish registered nurses [31], the response rate to the survey remained rather low. This may limit the generalizability of the findings to a larger research population. Second, we were able to use only 3 items from the scale measuring cognitive failures (WCFS), and the reliability of this measure (0.59) can be considered low. The WCFS has demonstrated high internal consistency for the whole scale and subscale level and the 3 items that were chosen for this study are the most indicative of the 3 components (attention, memory, function) of cognitive failure [20]. Third, although we controlled the analysis for gender and employment sector, we are aware

that some other variables may have contributed to stress related to information systems and cognitive failures as well (such as how long a person has used the current EHR system). Finally, the cross-sectional design did not allow the detection of causal relationships of the variables under study. The data used in this study was based on the first national survey of Finnish nurses gathered using the validated NuHISS [34]. A resurvey will be conducted in 2020, which will allow for further investigation of this topic.

Conclusions

Poor usability of EHRs can place a significant strain on the day-to-day work of nurses. This study suggests that cognitive performance, especially among young nurses, may be disturbed due to poor EHR usability. Young nurses need support and familiarization in many aspects of nursing during their first years in practice, and attention should be paid to providing them with appropriate support and training in the use of EHRs, which takes up a considerable amount of their working time. The results indicate that young nurses, who are typically believed to be fluent information technology users, may be burdened with poorly functioning information systems, possibly even more than their older colleagues are. System vendors have the primary responsibility to ensure the usability of their systems and to contribute to the quality of care and patient safety. It could be useful to investigate whether some usability factors are more critical than others. However, addressing the weaknesses of EHRs may be slow. In order to tackle the adverse consequences, it is important that employers provide adequate support for the right groups and that educational institutions provide students with adequate training in the use of EHRs. Further research should pay attention to the experiences of nurses of different ages and at different career stages in relation to the use of EHRs. It would also be useful to investigate the relationship between EHR education received as a student and early career stress and cognitive burden related to information systems. Finally, studies with longitudinal designs are needed to detect causal associations such as whether usability problems lead to cognitive failures.

Acknowledgments

This work was funded by the Strategic Research Council at the Academy of Finland (project 327145) and the Ministry of Social Affairs and Health (project 112241).

Authors' Contributions

AMK, KG, and TH conceptualized the study and developed the methodology. AMK and KG conducted the analysis. AMK wrote the original draft. All authors reviewed and edited the manuscript. TH provided supervision for the study.

Conflicts of Interest

None declared.

References

1. 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]
2. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011 May;4:47-55 [FREE Full text] [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
3. Vikkelsø S. Subtle redistribution of work, attention and risks: Electronic patient records and organisational consequences. Scandinavian Journal of Information Systems 2005;17(1):10 [FREE Full text]
4. Irizarry T, Barton AJ. A sociotechnical approach to successful electronic health record implementation: five best practices for clinical nurse specialists. Clin Nurse Spec 2013;27(6):283-285 [FREE Full text] [doi: 10.1097/NUR.0b013e3182a872e3] [Medline: 24107749]
5. Fleming NS, Culler SD, McCorkle R, Becker ER, Ballard DJ. The financial and nonfinancial costs of implementing electronic health records in primary care practices. Health Aff (Millwood) 2011 Mar;30(3):481-489. [doi: 10.1377/hlthaff.2010.0768] [Medline: 21383367]
6. Gephart SM, Bristol AA, Dye JL, Finley BA, Carrington JM. Validity and Reliability of a New Measure of Nursing Experience With Unintended Consequences of Electronic Health Records. Comput Inform Nurs 2016 Oct;34(10):436-447. [doi: 10.1097/CIN.0000000000000285] [Medline: 27551947]
7. Bristol AA, Nibbelink CW, Gephart SM, Carrington JM. Nurses' Use of Positive Deviance When Encountering Electronic Health Records-Related Unintended Consequences. Nurs Adm Q 2018;42(1):E1-E11. [doi: 10.1097/NAQ.0000000000000264] [Medline: 29194338]
8. Wisner K, Lyndon A, Chesla CA. The electronic health record's impact on nurses' cognitive work: An integrative review. Int J Nurs Stud 2019 Jun;94:74-84. [doi: 10.1016/j.ijnurstu.2019.03.003] [Medline: 30939418]
9. Belden JL, Grayson R, Barnes J. Defining and testing EMR usability: Principles and proposed methods of EMR usability evaluation and rating. Healthcare Information and Management Systems Society (HIMSS). 2009. URL: https://www.researchgate.net/publication/

277829258_Defining_and_Testing_EMR_Usability_Principles_and_Proposed_Methods_of_EMR_Usability_Evaluation_and_Rating [accessed 2020-10-27]
10. Viitanen J, Hyppönen H, Lääveri T, Vänskä J, Reponen J, Winblad I. National questionnaire study on clinical ICT systems proofs: physicians suffer from poor usability. Int J Med Inform 2011 Oct;80(10):708-725. [doi: 10.1016/j.ijmedinf.2011.06.010] [Medline: 21784701]
11. Kaipio J, Kuusisto A, Hyppönen H, Heponiemi T, Lääveri T. Physicians' and nurses' experiences on EHR usability: Comparison between the professional groups by employment sector and system brand. Int J Med Inform 2020 Feb;134:104018 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.104018] [Medline: 31835158]
12. Topaz M, Ronquillo C, Peltonen L, Pruinelli L, Sarmiento RF, Badger MK, et al. Nurse Informaticians Report Low Satisfaction and Multi-level Concerns with Electronic Health Records: Results from an International Survey. AMIA Annu Symp Proc 2016;2016:2016-2025 [FREE Full text] [Medline: 28269961]
13. Gephart S, Carrington JM, Finley B. A Systematic Review of Nurses' Experiences With Unintended Consequences When Using the Electronic Health Record. Nurs Adm Q 2015;39(4):345-356. [doi: 10.1097/NAQ.0000000000000119] [Medline: 26340247]
14. Koivunen M, Saranto K. Nursing professionals' experiences of the facilitators and barriers to the use of telehealth applications: a systematic review of qualitative studies. Scand J Caring Sci 2018 Mar;32(1):24-44. [doi: 10.1111/scs.12445] [Medline: 28771752]
15. Vehko T, Hyppönen H, Puttonen S, Kujala S, Ketola E, Tuukkanen J, et al. Experienced time pressure and stress: electronic health records usability and information technology competence play a role. BMC Med Inform Decis Mak 2019 Aug 14;19(1):160 [FREE Full text] [doi: 10.1186/s12911-019-0891-z] [Medline: 31412859]
16. Colligan L, Potts HW, Finn CT, Sinkin RA. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. Int J Med Inform 2015 Jul;84(7):469-476. [doi: 10.1016/j.ijmedinf.2015.03.003] [Medline: 25868807]
17. Saitwal H, Feng X, Walji M, Patel V, Zhang J. Assessing performance of an Electronic Health Record (EHR) using Cognitive Task Analysis. Int J Med Inform 2010 Jul;79(7):501-506. [doi: 10.1016/j.ijmedinf.2010.04.001] [Medline: 20452274]
18. Park Y, Kim SY. Impacts of Job Stress and Cognitive Failure on Patient Safety Incidents among Hospital Nurses. Saf Health Work 2013 Dec;4(4):210-215. [doi: 10.1016/j.shaw.2013.10.003] [Medline: 24422177]
19. Martin M. Cognitive failure: Everyday and laboratory performance. Bull Psychon Soc 2013 Oct 24;21(2):97-100. [doi: 10.3758/bf03329964]
20. Wallace JC, Chen G. Development and validation of a work-specific measure of cognitive failure: Implications for occupational safety. J Occup Organ Psychol 2005;78(4):615-632 [FREE Full text] [doi: 10.1348/096317905X37442]
21. Abbasi M, Zakerian A, Kolahdouzi M, Mehri A, Akbarzadeh A, Ebrahimi MH. Relationship between Work Ability Index and Cognitive Failure among Nurses. Electron Physician 2016 Mar;8(3):2136-2143 [FREE Full text] [doi: 10.19082/2136] [Medline: 27123223]
22. Harris DA, Haskell J, Cooper E, Crouse N, Gardner R. Estimating the association between burnout and electronic health record-related stress among advanced practice registered nurses. Appl Nurs Res 2018 Oct;43:36-41. [doi: 10.1016/j.apnr.2018.06.014] [Medline: 30220361]
23. Kowitlawakul Y. The technology acceptance model: predicting nurses' intention to use telemedicine technology (eICU). Comput Inform Nurs 2011 Jul;29(7):411-418. [doi: 10.1097/NCN.0b013e3181f9dd4a] [Medline: 20975536]
24. Khezri H, Abdekhoda M. Assessing nurses' informatics competency and identifying its related factors. Journal of Research in Nursing 2019 Apr 16;24(7):529-538. [doi: 10.1177/1744987119839453]
25. Ward MM, Vartak S, Schwichtenberg T, Wakefield DS. Nurses' perceptions of how clinical information system implementation affects workflow and patient care. Comput Inform Nurs 2011 Sep;29(9):502-511. [doi: 10.1097/NCN.0b013e31822b8798] [Medline: 21825972]
26. Heponiemi T, Hyppönen H, Vehko T, Kujala S, Aalto A, Vänskä J, et al. Finnish physicians' stress related to information systems keeps increasing: a longitudinal three-wave survey study. BMC Med Inform Decis Mak 2017 Oct 17;17(1):147 [FREE Full text] [doi: 10.1186/s12911-017-0545-y] [Medline: 29041971]
27. Heponiemi T, Kujala S, Vainiomäki S, Vehko T, Lääveri T, Vänskä J, et al. Usability Factors Associated With Physicians' Distress and Information System-Related Stress: Cross-Sectional Survey. JMIR Med Inform 2019 Nov 05;7(4):e13466 [FREE Full text] [doi: 10.2196/13466] [Medline: 31687938]
28. Babbott S, Manwell LB, Brown R, Montague E, Williams E, Schwartz M, et al. Electronic medical records and physician stress in primary care: results from the MEMO Study. J Am Med Inform Assoc 2014 Feb;21(e1):e100-e106. [doi: 10.1136/amiajnl-2013-001875] [Medline: 24005796]
29. Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM, American Medical Informatics Association. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc 2013 Jun;20(e1):e2-e8 [FREE Full text] [doi: 10.1136/amiajnl-2012-001458] [Medline: 23355463]

30. Holden R. Cognitive performance-altering effects of electronic medical records: An application of the human factors paradigm for patient safety. Cognition, Technology & Work 2011 Mar;13(1):11-29 [FREE Full text] [doi: 10.1007/s10111-010-0141-8] [Medline: 21479125]
31. Hyppönen H, Lääveri T, Hahtela N, Suutarla A, Sillanpää K, Kinnunen UM, et al. Smart systems for capable users? Nurses' experiences on patient information systems 2017. Finnish Journal of eHealth and eWelfare 2018;1(10):30-59. [doi: 10.23996/fjhw.65363]
32. Kinnunen U, Heponiemi T, Rajalahti E, Ahonen O, Korhonen T, Hyppönen H. Factors Related to Health Informatics Competencies for Nurses-Results of a National Electronic Health Record Survey. Comput Inform Nurs 2019 Aug;37(8):420-429. [doi: 10.1097/CIN.0000000000000511] [Medline: 30741730]
33. Reponen J, Kangas M, Hämäläinen P. Use of information and communications technology in Finnish health care in 2014. Current situation and trends. Helsinki: National Institute for Health and Welfare (THL); 2015. URL: https://www.julkari.fi/bitstream/handle/10024/126470/URN_ISBN_978-952-302-486-1.pdf [accessed 2020-10-27]
34. Hyppönen H, Kaipio J, Heponiemi T, Lääveri T, Aalto A, Vänskä J, et al. Developing the National Usability-Focused Health Information System Scale for Physicians: Validation Study. J Med Internet Res 2019 May 16;21(5):e12875 [FREE Full text] [doi: 10.2196/12875] [Medline: 31099336]
35. Heponiemi T, Hyppönen H, Kujala S, Aalto A, Vehko T, Vänskä J, et al. Predictors of physicians' stress related to information systems: a nine-year follow-up survey study. BMC Health Serv Res 2018 Apr 13;18(1):284 [FREE Full text] [doi: 10.1186/s12913-018-3094-x] [Medline: 29653530]
36. Kalakoski V, Selinheimo S, Valtonen T, Turunen J, Käpykangas S, Ylisassi H, et al. Effects of a cognitive ergonomics workplace intervention (CogErg) on cognitive strain and well-being: a cluster-randomized controlled trial. A study protocol. BMC Psychol 2020 Jan 02;8(1):1 [FREE Full text] [doi: 10.1186/s40359-019-0349-1] [Medline: 31898551]
37. Labrague L, McEnroe-Petitte D. Job stress in new nurses during the transition period: an integrative review. Int Nurs Rev 2018 Dec;65(4):491-504. [doi: 10.1111/inr.12425] [Medline: 29266201]
38. Halpin Y, Terry LM, Curzio J. A longitudinal, mixed methods investigation of newly qualified nurses' workplace stressors and stress experiences during transition. J Adv Nurs 2017 Nov;73(11):2577-2586. [doi: 10.1111/jan.13344] [Medline: 28543602]
39. Rudman A, Gustavsson JP. Early-career burnout among new graduate nurses: a prospective observational study of intra-individual change trajectories. Int J Nurs Stud 2011 Mar;48(3):292-306. [doi: 10.1016/j.ijnurstu.2010.07.012] [Medline: 20696427]
40. Deligkaris P, Panagopoulou E, Montgomery A, Maseura E. Job burnout and cognitive functioning: A systematic review. Work & stress 2014;28(2):107-123. [doi: 10.1080/02678373.2014.909545]
41. Salminen H. Turning the tide: Registered nurses' job withdrawal intentions in a Finnish university hospital. SA Journal of Human Resource Management 2012 Feb 17;10(2):1-11 [FREE Full text] [doi: 10.4102/sajhrm.v10i2.410]
42. Rudman A, Omne-Pontén M, Wallin L, Gustavsson PJ. Monitoring the newly qualified nurses in Sweden: the Longitudinal Analysis of Nursing Education (LANE) study. Hum Resour Health 2010 Apr 27;8:10 [FREE Full text] [doi: 10.1186/1478-4491-8-10] [Medline: 20423491]
43. Elfering A, Grebner S, Dudan A. Job characteristics in nursing and cognitive failure at work. Saf Health Work 2011 Jun;2(2):194-200 [FREE Full text] [doi: 10.5491/SHAW.2011.2.2.194] [Medline: 22953202]
44. Miller L, Stimely M, Matheny P, Pope M, McAtee R, Miller K. Novice nurse preparedness to effectively use electronic health records in acute care settings: Critical informatics knowledge and skill gaps. Online Journal of Nursing Informatics (OJNI) 2014;18(2) [FREE Full text]
45. Forman TM, Armor DA, Miller AS. A Review of Clinical Informatics Competencies in Nursing to Inform Best Practices in Education and Nurse Faculty Development. Nurs Educ Perspect 2020;41(1):E3-E7. [doi: 10.1097/01.NEP.0000000000000588] [Medline: 31860501]
46. Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: A systematic review. Health Policy 2018 Aug;122(8):827-836. [doi: 10.1016/j.healthpol.2018.05.014] [Medline: 29895467]
47. Gibson L, Sodeman W. Millennials and technology: Addressing the communication gap in education and practice. Organization Development Journal 2014;32(4):63-75 [FREE Full text]

Abbreviations

EHR: electronic health record
NuHISS: National Usability-Focused Health Information System Scale
WCFS: Workplace Cognitive Failure Scale

Edited by C Lovis; submitted 18.08.20; peer-reviewed by Q Do, J Tayaben; comments to author 26.09.20; revised version received 01.10.20; accepted 11.10.20; published 18.11.20.

Please cite as:
Kaihlanen AM, Gluschkoff K, Hyppönen H, Kaipio J, Puttonen S, Vehko T, Saranto K, Karhe L, Heponiemi T
The Associations of Electronic Health Record Usability and User Age With Stress and Cognitive Failures Among Finnish Registered Nurses: Cross-Sectional Study
JMIR Med Inform 2020;8(11):e23623
URL: http://medinform.jmir.org/2020/11/e23623/
doi:10.2196/23623
PMID:33206050

Original Paper

Using Ambient Assisted Living to Monitor Older Adults With Alzheimer Disease: Single-Case Study to Validate the Monitoring Report

Maxime Lussier[1,2], PsyD, PhD; Aline Aboujaoudé[1], MSc; Mélanie Couture[3,4], PhD; Maxim Moreau[5], PhD; Catherine Laliberté[6], MSc; Sylvain Giroux[6], PhD; Hélène Pigot[6], PhD; Sébastien Gaboury[7], PhD; Kévin Bouchard[7], PhD; Patricia Belchior[1,8], PhD; Carolina Bottari[2], PhD; Guy Paré[5], PhD; Charles Consel[9], PhD; Nathalie Bier[1,2], PhD

[1]Research Center of Institut universitaire de gériatrie de Montréal, Integrated Health and Social Services University Network for South-Central Montreal, Montreal, QC, Canada

[2]School of Rehabilitation, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[3]Integrated Health and Social Services University Network for West-Central Montreal, Université de Sherbrooke, Sherbrooke, QC, Canada

[4]Department of Psychology, Université de Sherbrooke, Sherbrooke, QC, Canada

[5]Research Chair in Digital Health, High Commercial Studies of Montreal, Montreal, QC, Canada

[6]Faculty of Sciences and Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, QC, Canada

[7]Department of Mathematics and Computer Science, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada

[8]School of Physical and Occupational Therapy, McGill University, Montreal, QC, Canada

[9]Bordeaux Institute of Technology & Inria, Bordeaux, France

Corresponding Author:
Maxime Lussier, PsyD, PhD
Research Center of Institut universitaire de gériatrie de Montréal
Integrated Health and Social Services University Network for South-Central Montreal
4545 chemin Queen-Mary
Montreal, QC, H3W 1W6
Canada
Phone: 1 514 340 3540
Email: lussier.maxime@gmail.com

Abstract

Background: Many older adults choose to live independently in their homes for as long as possible, despite psychosocial and medical conditions that compromise their independence in daily living and safety. Faced with unprecedented challenges in allocating resources, home care administrators are increasingly open to using monitoring technologies known as ambient assisted living (AAL) to better support care recipients. To be effective, these technologies should be able to report clinically relevant changes to support decision making at an individual level.

Objective: The aim of this study is to examine the concurrent validity of AAL monitoring reports and information gathered by care professionals using triangulation.

Methods: This longitudinal single-case study spans over 490 days of monitoring a 90-year-old woman with Alzheimer disease receiving support from local health care services. A clinical nurse in charge of her health and social care was interviewed 3 times during the project. Linear mixed models for repeated measures were used to analyze each daily activity (ie, sleep, outing activities, periods of low mobility, cooking-related activities, hygiene-related activities). Significant changes observed in data from monitoring reports were compared with information gathered by the care professional to explore concurrent validity.

Results: Over time, the monitoring reports showed evolving trends in the care recipient's daily activities. Significant activity changes occurred over time regarding sleep, outings, cooking, mobility, and hygiene-related activities. Although the nurse observed some trends, the monitoring reports highlighted information that the nurse had not yet identified. Most trends detected in the monitoring reports were consistent with the clinical information gathered by the nurse. In addition, the AAL system detected changes in daily trends following an intervention specific to meal preparation.

Conclusions: Overall, trends identified by AAL monitoring are consistent with clinical reports. They help answer the nurse's questions and help the nurse develop interventions to maintain the care recipient at home. These findings suggest the vast potential

of AAL technologies to support health care services and aging in place by providing valid and clinically relevant information over time regarding activities of daily living. Such data are essential when other sources yield incomplete information for decision making.

(JMIR Med Inform 2020;8(11):e20215) doi:10.2196/20215

KEYWORDS

activities of daily living; aging; Alzheimer disease; ambient assisted living; health care; technology assessment; health; remote sensing technology

Introduction

Background

Neurocognitive disorders affect 50 million individuals globally, with nearly 10 million new cases diagnosed each year [1]. Alzheimer disease (AD), the most common cause of dementia, is a disabling condition that has detrimental effects on memory, thinking abilities, behavior, and the ability to perform daily activities. By 2050, it is expected that one new case of AD will develop every 33 seconds, resulting in nearly 1 million new cases per year [2]. These numbers translate into an important global economic burden. Although the cost of providing medical and social care to individuals with dementia differs by country, the annual global cost of dementia in 2019 was estimated at US $1 trillion and is expected to double by 2030 [1]. As such, the World Health Organization recognizes it as a public health priority [3].

From the perspective of individual functional status, the speed and magnitude of decline varies, but the functional loss continuum begins with difficulties in completing complex instrumental tasks and continues until there is a complete loss of the ability to perform basic activities of daily living [4]. Instrumental tasks, including cooking, housekeeping, taking public transportation, and managing medication and finances, start declining in mild to moderate stages of the disease [4]. Basic activities of daily living—self-care activities such as eating, hygiene, grooming, and dressing—start declining in the moderate to severe stages of AD [4]. As such, throughout the progression of AD, older adults can become vulnerable to self-neglect.

Self-neglect is a behavioral condition that is characterized by the inability to sustain primary personal needs, including sufficient intake of food, personal hygiene, taking medication, and living safely [5]. Self-neglect predisposes older adults to devastating consequences such as multiple emergency department visits, maltreatment, nutritional deficiencies, nonadherence to medical treatment, and higher rates of morbidity, mortality, and nursing home placement [6-11]. The prevalence of self-neglect in older adults is high, ranging from 5% to 21%, on the basis of several factors [9], with the presence of cognitive impairments being the most important factor [12,13] even when the decline is mild [14-17]. The resulting decrease in independence in daily living has a significant impact on the individual's ability to stay at home, which causes an increased risk of institutionalization [18,19]. However, research has shown that older adults with neurological impairments prefer to live independently in their own homes even if they may be dependent on others for managing their daily lives [20]. Moreover, recent research has found that, even with cognitive impairment, living at home is more beneficial to older adults than relocating to a nursing home. For example, it was shown that older adults living at home experience a better quality of life, have better cognitive function, are less depressed, and are more socially active, with effects remaining even after stratifying for severity of dementia [21-23].

Ambient Assisted Living

Considering the growth of the aging population, the unprecedented economic pressures on the health and social care system this entails, and the benefits of staying at home for older adults, there is an urgent need to provide sufficient home care support to individuals with cognitive disorders.

In recent years, the use of ambient assisted living (AAL) systems has emerged as a way of promoting and extending aging in place. Implementing AAL systems involves deploying technologies (eg, sensors and actuators) in a home environment for the purpose of collecting continuous and real-time monitoring information about the environment (eg, temperature, humidity, and smoke in the home), the occupants (activities of daily living routines), and their health (eg, heart rate, body temperature, blood pressure, and blood oxygen level) [24]. AAL monitoring has many uses, such as detecting the occurrence of unusual or hazardous events (eg, a fall, cardiac arrest, or bradycardia) and alerting a dedicated resource person to provide immediate support and prevent dangerous situations. Another approach uses rich and reliable data from AAL monitoring to support clinical decision making regarding a care recipient's state [25-27]. For example, home care professionals have used home monitoring to adjust their care plan on the basis of clinical and home monitoring data [27]. For instance, they could add interventions if through monitoring they observe that a care recipient is sleeping too much or skipping meals. In addition, in the context of applying technological solutions in the care of seniors with AD, Kaye [28] proposed that AAL monitoring may be used to capture meaningful real-time changes in individuals' dementia trajectories and therefore plan preventive strategies. Considering the current limitations of functional assessments and the challenges associated with AD, Kaye suggested that AAL monitoring be used to improve health maintenance by offering "the opportunity to not only observe change in the person's usual environment but also to more frequently, and in some cases continuously, monitor a subject for salient change" [28]. Indeed, this could facilitate the development of timely strategies to maintain independence through later life or predict progression to disability.

Although novel, AAL monitoring has already shown promising outcomes. A recent literature review [26] concluded that the most promising technologies are those that monitor activities of daily living and detect falls and changes in health status. In 2016, a comprehensive review of the frameworks and sensors used in various AAL systems also showed that such technologies are often used to assess immediate safety risks to promote older adults living independently by alerting someone when a worrying situation arises [29]. However, these authors found that AAL systems are not often used for analytics and decision making, particularly for long-term care. For example, environmental factors such as temperature, humidity, motion, light, and contact could be analyzed to monitor health decline or the efficacy of an intervention.

Nevertheless, implementing AAL monitoring in real-life clinical settings presents several challenges. According to Peetom et al [26], to provide data effectively, it is essential that monitoring technology be based on algorithms that enable clinically relevant changes and situations to be detected without an overabundance of false alarms. Although AAL monitoring can predict cognitive decline among large cohorts of older adults [30], it remains to be shown whether it can support clinical decision making at an individual level (eg, for a care recipient). Clinicians have raised concerns about technology overloading them with information, especially irrelevant information [25], and generally clinicians have little office time to examine lengthy reports. Thus, it is important for them to be able to visualize collected data in a way that is intuitive and relevant for clinical decision making [31]. Other studies have used machine learning approaches to predict conditions such as cognitive decline [32-36]. For example, Dawadi et al [33] used statistical features (variances, autocorrelation, skewness, kurtosis, and change) of daily activity behavior (ie, total sensor events, cook duration, sleep duration) to train machine learning algorithms to predict the clinical assessment scores. With it, they achieved an accuracy of 72% in classifying cognitive assessment scores. Although these are promising and effective approaches, the process to which scores are calculated based on machine learning are at times described as a black box, owing to their lack of transparency [37]. By lack of transparency, we mean that it can be difficult for a clinician to grasp *why* the system has reached a given conclusion and, afterward, to explain or justify how these scores influenced their decision making (ie, understanding *why* the system has reached a given conclusion can be unfathomable). Transparency is critical for commercial, legal, and clinical applications because professionals must justify their decisions on the basis of tangible observations [38]. This finding was supported in preliminary focus groups conducted within the context of this study, as care professionals explicitly stated that they did not wish to be provided a conclusion or told what to do; instead, they wished to receive valid information that could be used to enhance their decision making [27].

As such, in this study, we provided a care professional with AAL monitoring reports that showed transparent information on their care recipients' daily routines (ie, time spent per day performing daily activities and standard deviation in time spent performing them). Statistical analyses applied to the collected data highlighted significant trends. This was done in a way that allowed the care professional to easily distinguish normal fluctuations from presumed significant changes in daily routines and to allow the care professional interpret these trends and take them into account in their decision-making process.

Objectives

This study seeks to provide essential insights into the clinical relevance of AAL use in real-life settings to support the delivery of home care services over time. Specifically, we seek to examine the concurrent validity of AAL monitoring reports and the well-established methods used by care professionals to gather information, given that any discrepancy between these 2 methods would lessen the perceived value of monitoring reports. To achieve this goal, a single-case study design is used to compare the information gathered regarding one care recipient with AD. A single-case approach is recommended when changes over time need to be assessed repeatedly [39]. Moreover, it allows for a more in-depth examination of consistency between events that would be difficult to translate into group means, as is the case with AAL monitoring reports.

Methods

This longitudinal single-case study is part of a larger project designed to understand how AAL monitoring can be successfully implemented in home care services and support the decision making of health and social care professionals in the public social and health care system [27]. The case in this study comprised one care recipient (*Lisette*) and the health and social care professional responsible for her home care support (referred to here as *the care professional*). Lisette's assigned case manager is a clinical nurse with a Bachelor's degree in nursing. She evaluates client health status and ensures that the nursing care and treatment plan is implemented for patients with complex health problems and provides care and treatment. As a case manager, her duties also include coordinating Lisette's specific needs for care and services and supervising her home support [40].

Recruitment

Health and social care professionals from the home care services of a local community health and social services center were invited to identify older adults who could benefit from AAL monitoring technology. To be included in the study, a care recipient had to be (1) receiving home care services due to a loss in functional autonomy related to a cognitive decline and (2) living alone. Care recipients were informed that AAL monitoring technology would be integrated into their home to monitor their daily routine (eg, sleeping and cooking). Care professionals were told that they would receive monthly reports on such activities to better understand the person's daily functioning. They would also be interviewed to better understand how they found the experience of using the monitoring data reports. The project was approved by the Aging and Neuroimaging Ethics Review Board of the South-Central Montreal Integrated University Health and Social Services Center (CER VN 16-17-22). All participants signed an informed consent form before taking part in the data collection process. Lisette was identified as a fitting care recipient for the study by her care professional in January 2017.

Data Collection

The AAL monitoring technology was installed in Lisette's home on February 9, 2017. Her participation in the project ended a year and a half later, after her hospitalization and transfer to a long-term care facility, on July 6, 2018. Her data set was collected through a monitoring period that extended over 490 days, using 25 sensors and comprising approximately 617,000 logs. The sensors and algorithms used to create the monitoring reports are described later.

The monitoring reports were triangulated with multiple sources of data to examine their concurrent validity: medical file, verbatim of interviews with the care professional, emails, and memos of telephone exchanges (approximately 20) with the care professional. The care professional was interviewed on January 24, 2017 (before any monitoring report was sent); June 14, 2017; October 25, 2017; and October 26, 2018 (after Lisette's hospitalization). These interviews were conducted by a researcher specializing in qualitative research (MC) and triangulated with the monitoring report outcomes [27]. The final interview took place 3 months after Lisette's hospitalization due to uncertainty about Lisette's transfer during that period.

Case Description

Lisette was selected as a representative case for this study for several reasons. First, she was an older woman with AD who wished to remain in the apartment in which she had been living for several years for as long as possible. Second, the extensive data set collected through a 490-day monitoring period allowed for a rich analysis. Third, the monitoring began when the care professional became concerned about Lisette's ability to live independently and continued until her transfer due to the deterioration of her condition. Therefore, we were confident that documenting this individual during this particular period would encompass significant changes in behavior, from beginning to end.

Lisette, a widowed older woman, was 91 years old at the time of her recruitment and had been living in the same one-bedroom apartment for the last 14 years. Her apartment was in a residence for independent seniors. Her son was her main caregiver and the primary contact for health care providers in case of an emergency. In 2015, she was diagnosed with AD, but a vascular etiology was also considered. Despite the diagnosis, she wished to continue living in her apartment, with assistance, and developed a social network with her neighbors. Although her son was considering whether Lisette should be transferred to a facility with a higher level of care, her daughter did not consider it necessary; thus, they had conflicting points of view about their mother's level of independence. Their difference of opinion made it difficult for the care professional to gather reliable information to guide her own clinical decision making. Moreover, Lisette seemed to have limited awareness of her cognitive difficulties and their impact on daily living. Her son and the care professional mentioned that she was often unreliable when asked about her everyday routine and recent events.

In this context, Lisette's care professional wished to acquire objective and reliable information regarding Lisette's routine. This would enable her to better assess the safety risks related to maintaining Lisette in her home and to confirm or refute her clinical hypothesis before developing a comprehensive intervention plan. The main issues of concern identified by the care professional at the outset of the study were as follows: (1) malnutrition (because of food left untouched in the refrigerator over long periods of time and low body weight), (2) hazardous use of the stove (safety concerns expressed by the landlord), and (3) poor personal hygiene (slow shrinkage of the bar of soap observed by the care professional).

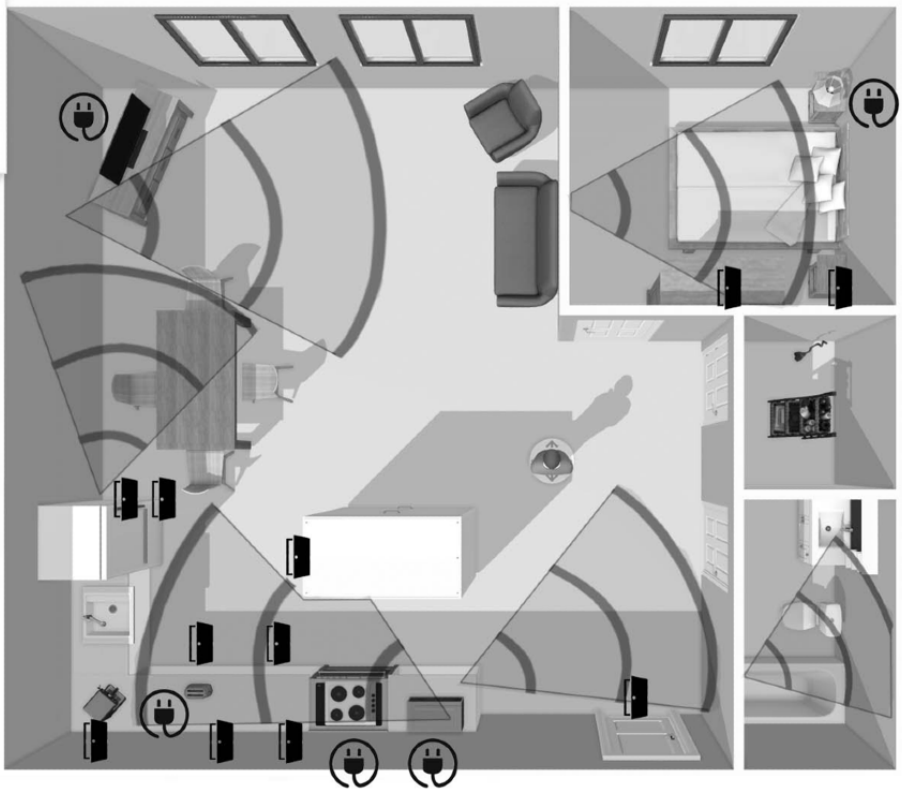
The monitoring period began on February 9, 2017, and ended on July 6, 2018. During this period, the care professional observed that Lisette's AD symptoms progressed. During follow-up clinical visits, Lisette's memory and orientation declined noticeably and continuously from December 2018 onward. Monitoring continued until Lisette experienced a serious episode of confusion and panic, calling her son because she was looking for her late husband. Following that episode, she was hospitalized on July 6, 2018, and shortly thereafter was moved by her family to a private home for older adults, ending the monitoring.

Sensors and Algorithms

Three different types of sensors were integrated into the home environment: passive infrared (PIR) sensors (Everspring HSP02), magnetic contact sensors (Everspring HSM02), and smart electric switches (Aeotec ZW078 and ZW096). An attempt to integrate water sensors was unsuccessful due to rapid corrosion of the sensors when submerged in water and difficulties mounting them in a humid environment. Wireless sensors were used because they were easy to install in a real-life setting and could be moved to a new apartment if needed; these were considered important characteristics for decision makers in the context of a public health care system.

One PIR sensor per room was installed in the kitchen, dining room, and living room; at the entrance; and in the bathroom. Two were installed in the bedroom: one aimed toward the bed and another directed toward the door by which to exit the room. Electric sensors were connected to a television, a bed lamp, a microwave, a toaster, and a stove. Contact sensors were installed on the entrance door, on 2 dresser drawers in the bedroom, and on 2 food storage cabinets, the refrigerator, freezer, oven, utensil drawer, and 4 cupboards in the kitchen. Figure 1 illustrates a map of sensor deployment in Lisette's apartment.

Figure 1. Sketch of the placement of sensors in Lisette's apartment. Conic area: passive infrared motion sensors; door symbol: magnetic contact sensors; electric plug symbol: smart electric switches.



The selection and placement of wireless sensors were customized to specifically focus on daily activities that were relevant to Lisette's care professional, namely, sleep habits (sleep), exiting the apartment (outing), periods of prolonged inactivity (low mobility), cooking-related activities (cooking), and hygiene-related activities performed in the bathroom (hygiene). Algorithms were built around various assumptions about these different activities, as previously described in a study by Lussier et al [27]. First, *room occupation* was recognized as follows: the occupant was considered to be in one room for as long as he or she was not detected in another room or outside the apartment. *Sleep* was identified if the occupant spent more than 20 min in the bedroom without interacting with any sensors other than the PIR directed at the bed. *Outing* was recognized if the following sequence of events occurred: (1) closing the entrance door, (2) no PIR activity in the home for more than 5 min, and (3) opening the entrance door. *Low mobility* was recognized if no PIR sensors and no contact sensors were triggered over a 15-min period and the occupant was not recognized as resting (in the bedroom) or being out. With respect to *cooking* activities (ie, meal preparation, washing the dishes, storing groceries), the assumption was that the occupant was cooking if several sensors placed in the kitchen were triggered over a short period. More precisely, for each 15-min period (ie, 3 PM to 3:15 PM, 3:01 PM to 3:16 PM, etc), a cooking score was established on the basis of the frequency and diversity of sensors triggered in the kitchen. If the score was 2 SDs higher than the average for that occupant, they were then recognized as cooking. This approach is similar to that used by Rantz et al [41] and was done to report only on activities that were significant in the context of the occupant's daily routine. Finally, given that water sensors could not be successfully installed on the sink and bathtub faucets, the recognition of *hygiene* (ie, brushing teeth, showering, going to the toilet, washing hands) relied exclusively on prolonged presence in the bathroom. Similar to the cooking activity, a score was calculated on the basis of the number of minutes spent in the bathroom per 15-min window. The score then had to be 2 SDs higher than the average for that occupant to be recognized as engaging in a period of bathroom hygiene. It is important to note that it was impossible to verify whether the occupant or someone else (eg, the caregiver) was performing the actions. The care professional was aware of this limitation. See Textbox 1 for an overview of detailed algorithms.

Textbox 1. Overview of daily activity algorithms.

```
##MAIN ALGORITHM
While(TRUE)
  Map(RoomName,Presence)=IndoorPIRMotionSensorList ()
  ROOMOCCUPATION=Map(RoomName,Presence)
  Sleeping=Event (Sleep(ROOMOCCUPATION))
  Outing=Event(Outings (ROOMOCCUPATION))
  LowMobility=Event(LMobility(ROOMOCCUPATION))
  Cooking=Event (Cook(ROOMOCCUPATION, standardDeviations(cookingScore)))
  Hygiene=Event (BathroomAct(ROOMOCCUPATION, standardDeviations(hygieneScore)))
End
MIN_OUTING_TIME=5 minutes
MIN_SLEEPING_TIME=20 minutes
MIN_LOWMOBILITY_TIME=20 minutes
##SUBFUNCTION ALGORITHM
##Sleep
IF(ROOMOCCUPATION in [Bedroom]>, MIN_SLEEPING_TIME)
  IndoorPIRMotionSensorLastTrigger([Bedhead] ,Duration))
##Outings
IF (ROOMOCCUPATION in [Entrance]]>MIN_OUTING_TIME)
ClosingEntranceDoor(True)
  IndoorPIRMotionSensorLastTrigger([],Duration)
##LMobility
IF(NOT(sleeping) & NOT(outing) & MagneticContactSensorLastTrigger>MIN_LOWMOBILITY_TIME & RoomOccupationLastChange>MIN_LOWMOBILITY_TIME)
IndoorPIRMotionSensorLastTrigger([] ,Duration))
##BathroomAct
IF(ROOMOCCUPATION in [Bathroom])
  BathroomSensorsUsed[IndoorPIRMotionSensorLastTrigger([Bathroom], MAX_ACTIVITY_TIME)]
  hygieneScore = FrequencyOfUse (BathroomSensorsUsed)
##Cook
IF(ROOMOCCUPATION in [Kitchen, Dining])
  KitchenSensorsUsed[IndoorPIRMotionSensorLastTrigger([Kitchen, DiningRoom], MAX_ACTIVITY_TIME)] +
  KitchenSensorsUsed[MagneticContactSensorLastTrigger ([Kitchen, DiningRoom], MAX_ACTIVITY_TIME)] +
  KitchenSensorsUsed[ElectricalMeasurementSensor ([Kitchen, DiningRoom], MAX_ACTIVITY_TIME))]
cookingScore = FrequencyOfUse(KitchenSensorsUsed)
```

Monitoring Reports and Statistical Analysis

Each month postbaseline, the care professional was presented with a monitoring report sent via email. Reports were sent monthly because this was considered frequent enough by the care professionals. Monitoring reports were divided into 2 sections. The first section shows features from the previous month (eg, the time of day when daily activities were most likely to occur, a pie chart of the different room occupancy averages within the home; see Lussier et al [27] for an example). The second section, which is the main focus of this paper, shows the monthly evolution of the average time per day spent performing each of the 5 daily activities. It included the evolution of the frequency and average duration of stove and microwave use as well.

To determine which lifestyle changes should be highlighted as significant to the care professional, linear mixed models for repeated measures were used for each activity (ie, sleep, outing activities, periods of low mobility, cooking-related activities, and hygiene-related activities). Outcomes were expressed as

the daily duration of time spent performing these activities. In Lisette's case, 490 continuous days of monitoring were divided into 14 months. Although the term *month* will be used for the sake of simplicity, 35-day periods were used instead of calendar months. This was done so that each month contained the same number of days, with 5 occurrences of each day of the week (ie, Monday to Sunday). This further allowed us to control for any discrepancy in weekly routines (eg, having a dance course every Sunday afternoon). Fixed factors were defined as months and weekdays. The covariance structure selected was compound symmetry for all the tested responses and was determined according to the Akaike information criterion. For outcome, *the marginal means of each month were compared with the initial baseline month. P* values were not adjusted for multiple comparisons as each observation was compared with the baseline observation. This was also done to preserve the statistical power of this exploratory experiment, as we did 13 comparisons, adjusting the *P* value accordingly was considered too restrictive. All analyses were conducted with an α threshold of .05, using IBM SPSS Statistics version 25.

For each activity, 2 sets of analyses were performed. First, we examined statistically significant overall trends, as a trend could correlate with cognitive or health decline. Second, we compared the baseline report (ie, the first month of monitoring) with each following month to determine if, and which, months significantly differed from the baseline. On the one hand, if a general trend was observed, this could be used to help determine at which month a trend reached significance. On the other hand, this could be used to identify irregular months that did not correspond to any overall trend. This could further be useful to pinpoint important but nonpermanent changes in the daily routine for the care professional to explore. It is also important to note that monitoring reports were part of an ongoing process; therefore, what could initially appear to be an irregular month could in fact become part of an ongoing trend as months passed by.

Concurrent Validity

To examine the concurrent validity of AAL monitoring reports for each of the main activities monitored (sleep, outing, low activity, cooking activities, and hygiene), we first used statistical analyses and values for changes in daily activities that were highlighted in the monitoring reports. We then explored whether any of the significant changes from the monitoring data could be linked with information gathered by the care professional. This information was extracted from interviews, emails, and memos exchanged with the care professional and information from the care recipient's medical file.

We assumed that significant changes calculated from the monitoring data would be coherent with real-life information, thereby validating the potential of this approach in a clinical setting. Moreover, because the care professional may lack continuous and reliable information, we expected that monitoring reports would detect changes in activities of daily living that the care professional would not have been aware of.

Results

For each of the main activities monitored (sleep, outing, low activity, cooking activities, and hygiene), the results are provided in 2 sections. The first section presents statistical analyses and values for significant changes in the monitoring reports (see Table 1 for means and 95% CI of repeated measurements). The second section compares each significant change with information gathered by the care professional to explore concurrent validity. For each value, the mean and SD are presented.

Table 1. Means and 95% CI of repeated measurements for sleeping outings, cooking activities, hygiene, and low mobility during the monitoring period.

Month	Sleep, mean (95% CI)	Outings, mean (95% CI)	Cooking activities, mean (95% CI)	Bathroom usage, mean (95% CI)	Low mobility, mean (95% CI)
1	8.36 (7.93 to 8.80)	3.11 (2.49 to 3.72)	1.50 (1.28 to 1.72)	1.96 (1.71 to 2.20)	3.96 (3.18 to 4.74)
2	8.53 (8.09 to 8.97)	3.65 (3.04 to 4.27)	1.24 (1.02 to 1.46)	2.22 (1.92 to 2.53)	3.88 (3.10 to 4.66)
3	8.66 (8.24 to 9.10)	3.15 (2.54 to 3.77)	1.17 (0.95 to 1.39)[a]	2.07 (1.79 to 2.35)	3.74 (2.96 to 4.52)
4	9.08 (8.60 to 9.57)[a]	2.73 (2.06 to 3.41)	1.02 (0.78 to 1.27)[a]	1.66 (1.39 to 1.93)	3.93 (3.06 to 4.80)
5	8.71 (8.27 to 9.15)	2.54 (1.93 to 3.16)	0.77 (0.55 to 0.99)[a]	1.66 (1.41 to 1.90)	4.12 (3.31 to 4.92)
6	9.37 (8.88 to 9.85)[a]	1.22 (0.60 to 1.85)[a]	0.73 (0.49 to 0.97)[a]	1.76 (1.51 to 2.01)	4.63 (3.76 to 5.51)
7	8.96 (8.48 to 9.43)	1.40 (0.76 to 2.05)[a]	0.80 (0.57 to 1.02)[a]	1.65 (1.40 to 1.91)	5.94 (5.11 to 6.77)[a]
8	9.42 (8.96 to 9.88)[a]	1.40 (0.78 to 2.01)[a]	1.10 (0.88 to 1.32)[a]	1.84 (1.60 to 2.09)	4.08 (3.25 to 4.91)
9	9.90 (9.42 to 10.39)[a]	0.57 (−0.06 to 1.21)[a]	1.03 (0.80 to 1.26)[a]	1.87 (1.62 to 2.13)	5.38 (4.55 to 6.21)[a]
10	9.77 (9.30 to 10.24)[a]	1.56 (0.94 to 2.18)[a]	1.14 (0.91 to 1.36)[a]	1.47 (1.22 to 1.72)[a]	4.21 (3.37 to 5.06)
11	9.74 (9.30 to 10.19)[a]	2.05 (1.44 to 2.67)[a]	1.71 (1.49 to 1.93)	1.62 (1.37 to 1.86)	3.83 (3.03 to 4.62)
12	9.72 (9.24 to 10.20)[a]	1.85 (1.24 to 2.47)[a]	1.76 (1.54 to 1.98)	1.72 (1.47 to 1.96)	4.84 (4.06 to 5.62)
13	10.11 (9.65 to 10.57)[a]	1.17 (0.53 to 1.80)[a]	1.23 (1.00 to 1.46)	1.56 (1.29 to 1.83)[a]	4.29 (3.49 to 5.10)
14	10.18 (9.74 to 10.62)[a]	0.95 (0.33 to 1.56)[a]	1.05 (0.83 to 1.27)[a]	1.36 (1.12 to 1.61)[a]	3.43 (2.64 to 4.21)

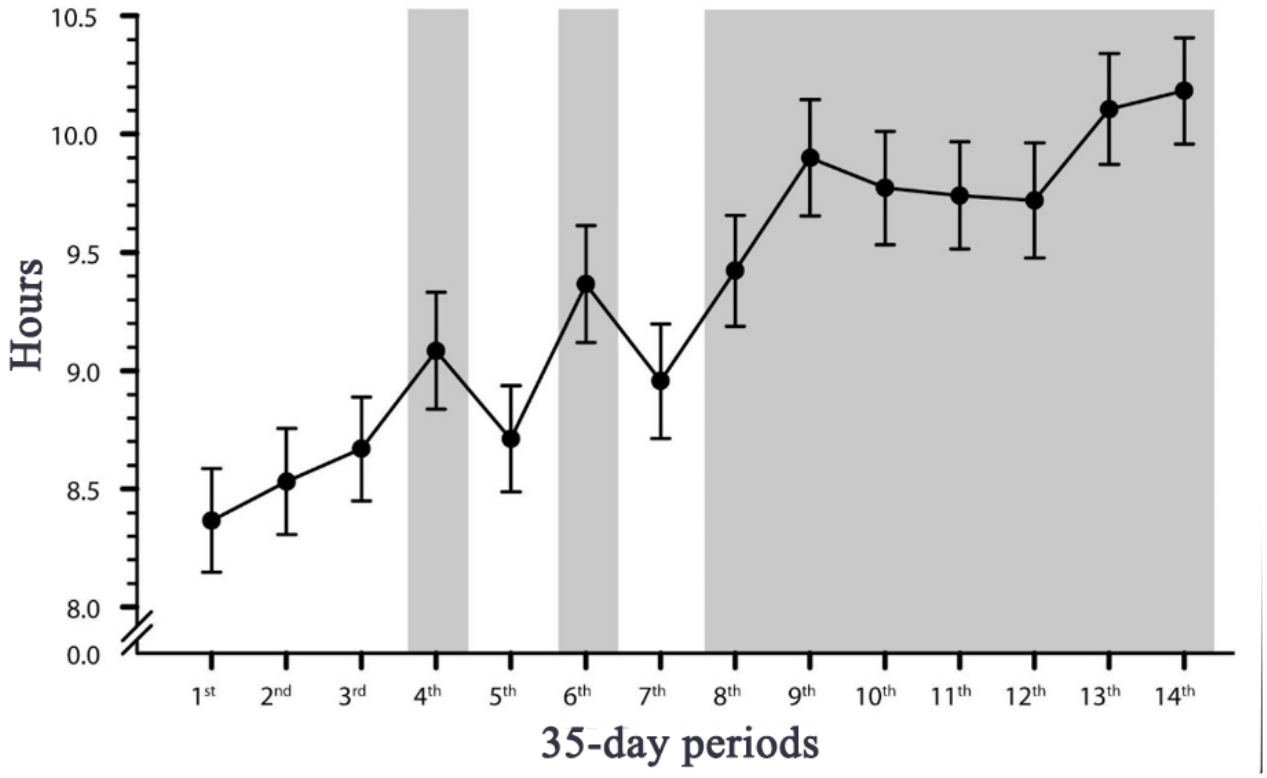
[a]Statistically significant (P<.05) changes compared with the first period.

Sleep Habits

Monitoring Report

Over the course of monitoring, Lisette gradually spent more time resting (month effect: $F_{13, 397.98}$=7.05; P<.001). This trend became steady and significant from the eighth month onward. In the last month, sleep was detected for 10.18 (SD 1.64) hours, which represents a 22% increase in resting time when compared with the baseline (8.37, SD 1.27 hours; Figure 2). In addition, the monitoring data suggested that she more frequently woke up later in the morning toward the end of the monitoring period. The data showed that she woke up between 7:25 AM and 7:56 AM during the first month but between 7:33 AM and 10:30 AM over the last month of monitoring. There were no indications that she woke up more frequently during the night.

Figure 2. Evolution of estimated means (SE) for hours of sleep detected during the monitoring period. The gray zones highlight statistically significant (P<.05) periods when compared with the first period.



Clinical Observation

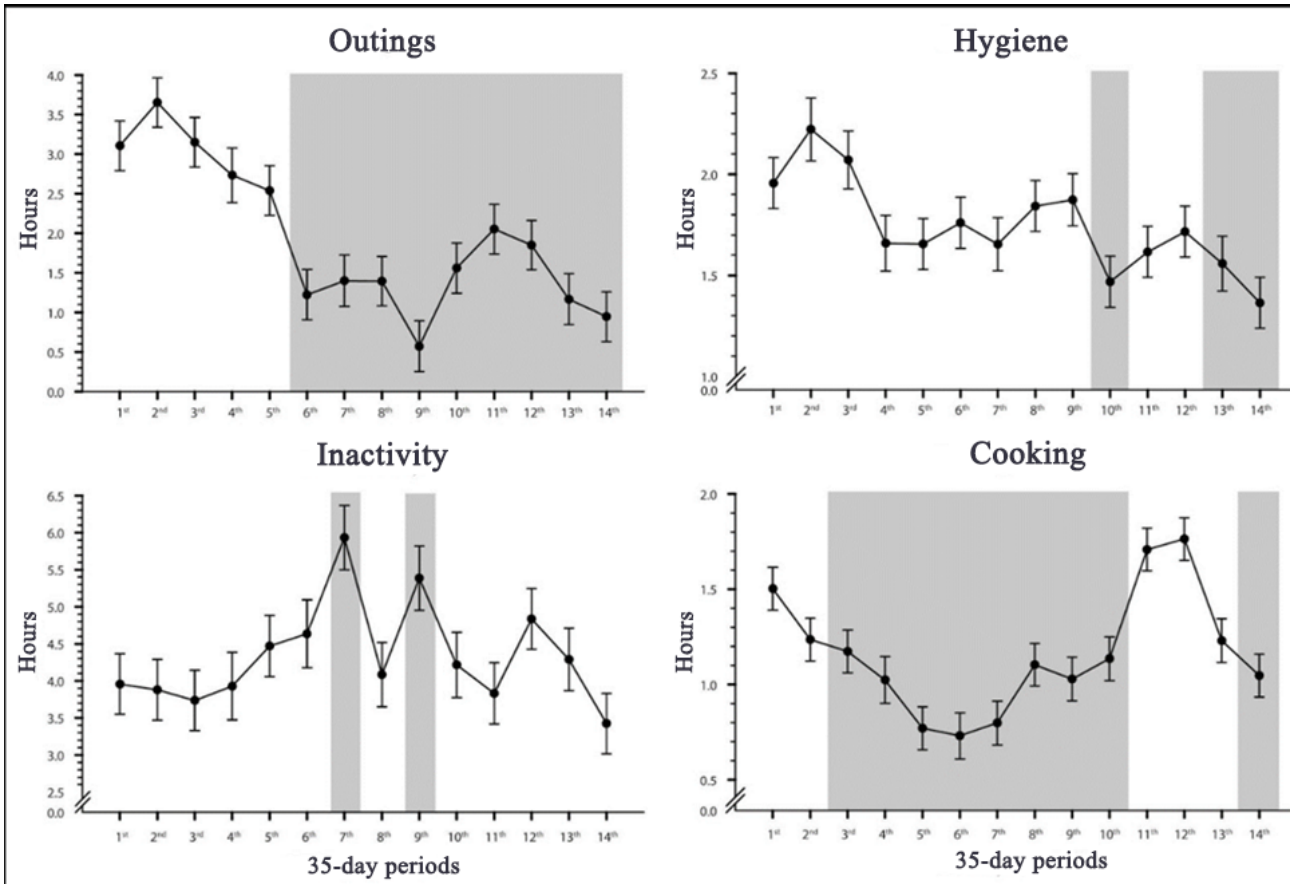
Interestingly, Lisette did not mention to the clinician that she was feeling more tired or sleeping longer than before. However, the personal care assistant responsible for giving her medication each morning (a different health professional, not the one in charge of coordination and treatment plan) noted that she had to wake her up more often when visiting.

Outings

Monitoring Report

Overall, Lisette gradually went outside for shorter periods (month effect: $F_{13,\ 443.00}$=8.70; P<.001; see Figure 3, top-left). This change became steady and significant from the sixth month onward. The reports highlight that Lisette spent 68% less time outside when compared with the baseline (3.10, SD 2.80 hours) and to the last month (0.97, SD 0.90 hours). Another noticeable decline in outings was observed specifically in December (0.35, SD 0.23 hours).

Figure 3. Evolution of estimated means (SE) for outings, bathroom usage, low mobility, and cooking activities during the monitoring period. The gray zones highlight statistically significant (P<.05) periods when compared with the first period.



Clinical Observation

Trends in outings were corroborated by Lisette, who mentioned to the care professional that she had lost interest in the social activities held at her building because she said she disliked the new hosts in charge of the activities. She also mentioned that she had stopped attending the dance activities on weekends because she did not appreciate the change in the style of music played. There was no way for the clinician to verify that information. As for the decline observed for the month of December, it was noted in her medical record that Lisette had influenza in December and was quite incapacitated by it, which would most likely account for her staying at home.

Cooking Activities

Monitoring Report

Over time, a significant decline in cooking-related activities was observed (month effect: $F_{13, 424.81}$=7.83; P<.001; see Figure 3, top-right). However, the time spent on these activities did not follow a steady decline. A first significant decline was observed in the third month, and the decline continued gradually up to the sixth month, the lowest point reached (42.24, SD 26.40 m), representing a 53% decline when compared with the baseline (90.00, SD 50.00 m). Cooking activities slowly increased over the 11th to 13th month, reaching a value comparable with that of the first month. Finally, cooking activities started decreasing again during the last month (62.40, SD 30.61 m), a 31% decline compared with the first month.

The stove was used on about 16% of days during the monitoring period (the oven with or without the burners, 13%, and only the burners, 3%). On days the stove was used, burners were used for an average of 10 (SD 11) min and the oven for 15 (SD 10) min. Only one instance of dangerous stove use was identified: Lisette left the apartment while the stove was switched on and then came back but only turned it off the next morning. The microwave was used on 32% of days for an average of 6 (SD 6) min per day. On one occasion, the microwave was in use for 20 min nonstop, which was a clear outlier for Lisette. Other than that, no unusual behaviors were detected.

Clinical Observation

The patterns related to cooking activities were supported by real-life information. Notably, when the care professional received monthly reports illustrating raw data, she noticed a change in pattern. Considering her concern that Lisette was at risk of malnutrition, she used that information to discuss Lisette's eating habits with her son during the summer (exact date unknown, but between the fifth and seventh month). This conversation led to a change in the caregiver's behaviors: now more aware that his mother preferred homemade meals to ready or frozen meals, and to increase meal intake, the son said he would prepare more homemade meals for her. He and his wife would also come more often to cook for her and put leftovers in the refrigerator. This decision correlated with increased detection of cooking-related activities in Lisette's home for several months after this exchange. Therefore, the increase in time spent cooking at that time is a result of the compensatory

behaviors by family members (ie, family cooking at Lisette's home and Lisette eating more regularly because the food she liked was easily available). However, this compensatory behavior either did not last or was not sufficient support, as Lisette's cooking-related activities again declined in the last 2 months before her hospitalization and the end of monitoring.

Overall, the infrequent use of the microwave and stove was consistent with the care professional's suspicions that Lisette might not be eating hot meals daily (as indicated by the accumulation of prepared meals in the refrigerator). Concerning the dangerous use of the stove, a member of the research team contacted Lisette shortly thereafter to verify if the sensor was defective. She said she forgot the burner on and burned her fudge during the night. The care professional was informed of the event. As for the microwave, the care professional questioned Lisette about this incident a couple of weeks later. She answered that she probably meant to set 2 min but added a *0*, resulting in 20 min instead.

Hygiene

Monitoring Report

The monitoring results suggest that hygiene-related activities declined significantly over time (month effect: $F_{13, 40658}$=2.94; P<.001), reaching a 30% decline in time spent in the bathroom when comparing the baseline (117.36, SD 56.80 m) with the last month (82.98, SD 25.80 m; see Figure 3, bottom-left). This trend first became significant during the 10th month but remained steadily significant only from the 13th month onward. Activation of the clothing drawers' sensors remained stable over time.

Clinical Observation

As changes in hygiene were detected late in the study, the care professional did not have many opportunities to gather information on this aspect before Lisette was hospitalized and transferred. However, she mentioned that a decline in hygiene would not be surprising, considering the rapid decline of her cognitive state in the month preceding her transfer. Moreover, post hoc analyses suggest that the longest bathroom hygiene periods tended to occur before outings. Therefore, a decline in hygiene-related activities would be consistent with Lisette's confirmed abandonment of some social activities.

Low Mobility

Monitoring Report

The average period of low mobility remained quite stable over time (month effect: nonsignificant.; see Figure 3, bottom-right). The only 2 significant increases in periods of low mobility occurred in the seventh month (5.27, SD 3.21 hours; P<.001) and the ninth month (4.69, SD 1.65 hours; P=.02) compared with baseline (3.96, SD 1.98 hours).

Clinical Observation

Although there was no clear explanation available for the increase in inactivity in the seventh month, Lisette's medical record reported that she was sick with influenza during the ninth month, which could explain the fewer outings and more inactivity being detected in the home.

Discussion

Principal Findings

This study describes a longitudinal single-case study in a Canadian public home care setting. The objective was to examine the concurrent validity of AAL monitoring reports and care professional descriptions of real-life changes in activities of daily living experienced by an older adult diagnosed with AD (ie, Lisette). Lisette's care professional received monthly monitoring reports of sleep, low mobility, outings, cooking, and hygiene-related activities. In the monitoring reports, algorithms and linear mixed models for repeated measures were used to highlight significant changes in Lisette's daily activities. Highlights from the reports were then juxtaposed to information gathered by the care professional to determine if they concurred. Lisette's health and life events were gathered from her medical file and from interviews, emails, and telephone exchanges with her care professional.

As expected, statistically significant changes in daily routine were detected over the 490 days of monitoring. Through interviews conducted with the care professional, it was possible to conclude that monitoring report trends were consistent with the clinical information collected when staff visited the care recipient. In fact, a priori interrogations and observations made by the care professional were indeed reflected in the monitoring reports. For instance, the care professionals believed that Lisette was not cooking complex meals, and this was later confirmed by the monitoring report. In addition, the outcome of subsequent interventions was observable in the monitoring report. This occurred when the care professional invited the family to participate more in meal preparation, which led to a detectable change in routine. Interestingly, in addition to validating the initial hypotheses, the monitoring reports drew attention to certain unforeseen or unexpected changes that were then triangulated with other information gathered by the clinician. On occasion, this led to discussions with the care recipient. For example, when there was a temporary decline in outings and an increase in low-mobility periods during the month of December, the care professional asked Lisette about it. She then learned that Lisette had caught the flu that month, which was supported by information in her medical file. Another example is the gradual increase in the time spent sleeping, which Lisette was not aware of (or denied) but was corroborated by a personal care assistant. As such, by combining information from the monitoring report with comments made by another care worker, the care professional was able to gain a better understanding of Lisette's situation that would have otherwise been overlooked, unattainable, or ambiguous.

Comparison With Prior Work

This study is innovative because monitoring reports were designed in collaboration with care professionals to specifically address their requirements and the care professional's need for information that would enable the best client support possible were carefully examined. Moreover, this technology was implemented in concert with the head of service to be included harmoniously with actual current services from the public health care system. Importantly, the evolution of daily routine observed

in this case study was highly consistent with the current literature on daily activities in AD. For instance, sleep disturbance is prevalent and predictive of cognitive decline in older adults and in those with neurocognitive disorders [42]. Three studies using AAL monitoring technology to monitor sleep found that sleep quality and sleep hygiene measures were related to mild cognitive disorders in older adults [43-45]. More precisely, it is suggested that AD is not associated with more sleeping time but with less sleep efficiency (ie, lower percentage of time in bed spent asleep) [46]. With the sensors used in this case study, it was not possible to distinguish the time spent sleeping from the time spent lying down in bed. Nonetheless, although Lisette did not report any change in her sleep routine when asked, it is possible to assume that her sleep quality decreased because she needed to spend more time lying down in her bedroom as the disease progressed. The decrease in time spent outside the home was partially explained by Lisette, who mentioned withdrawing from social activities. She justified this behavior by referring to different changes in the way the activities were held. However, this behavioral evolution is also consistent with several studies showing that older adults with cognitive impairment reduce their engagement in social activities as their cognition declines [47-50]. Anosognosia (ie, a lack of awareness of deficit) is often observed early in the course of AD [51], so it is possible that Lisette withdrew from social activities because of cognitive decline but without being aware of that change. It is also at this stage that agitation, confusion, and distress episodes, such as the one Lisette experienced just before being moved out of her apartment, occur more frequently [52]. Therefore, anxiety and distress might have kept her from going out. Finally, the decline in hygiene occurring last in the timeline is also consistent with the literature. Indeed, while a decline in the ability to perform basic activities of daily life is minimal in the early stages of AD, moderate stages bring incipient difficulties with this sphere [53]. For example, at this stage, the care recipient should be able to shower alone but may forget to do so regularly. Therefore, the decline in hygiene was consistent within the context of progressing dementia and could be considered an indicator of worsening symptoms.

In comparison with other similar studies on AAL and clinical reasoning, our study is also in line with the study by Rantz et al [41]. In their study, Rantz et al [41] showed that monitoring the increase in activity level in the bathroom could support the detection of early signs of urinary tract infection by a nurse. To our knowledge, our study is however the first to document the use of daily living monitoring over a long period, in relation to other clinical data to support the clinical reasoning of health care providers. Monitoring data related to activities of daily living have great potential to support the work of home care providers via telehealth modalities. Faced with unprecedented challenges in allocating resources, in urban and remote rural areas, vast countries such as Canada could use AAL tools to better allocate services to the right person, at the right moment, ensuring more equitable and sustainable use of public health care capacities [54].

Limitations

Although the results of this case study are promising, further replications among care recipients with similar medical conditions and in a variety of environments are necessary. The limitations of the monitoring technology used must also be addressed. First, it was impossible to accurately determine whether more than one occupant was in the home at the same time. Therefore, visitors' activities were averaged in the data reports. The multioccupant dilemma must be further explored to develop a simple but accurate solution. This solution should not require wearable sensors, as poor compliance has been reported with this technique in older adults with cognitive deficits [55]. Nevertheless, this study focused on the question, "Has the activity been done?" rather than "Who is doing the activity?" because our participant was living alone. From the perspective of the care professional using the monitoring reports to decide if any additional services are needed or not, the main interest is in knowing that the activity is being carried out regularly, regardless of who has done it; a decision taking into account the support given by the caregivers, not in abstraction of it. Furthermore, care professionals are particularly interested in AAL monitoring technology for care recipients who live alone with minimal caregiver support as social isolation is a major risk factor for security and reliable sources of information are often lacking in this context [27]. Finally, monitoring reports are not a substitute for a functional, performance-based evaluation and should not be used.

Conclusions

Although living at home presents several benefits for older adults [4,22,23], dealing with the increase in the incidence of self-neglect also poses great challenges for the health care system. This increases the demand for better resource allocation and innovative strategies to attenuate the estimated impact on health care expenditures. With the technology boom of the past decades, AAL monitoring systems are among the most promising tools to support the outcomes of individuals whose cognitive declines limit their effective participation in daily activities. AAL may improve the ability of older adults to cope at home and handle tasks and needs, which are the basis of the well-known aging-in-place design for living. Moreover, global efforts to cope with the COVID-19 pandemic of 2020 have acutely highlighted that it is crucial to have a system of health and social services that allow for remote monitoring of fragile older adults living alone and isolated under conditions of social distancing [56].

In this paper, we showed how an AAL monitoring system, using customized wireless sensors, was relevant during the assessment of home care services for an older woman with AD at risk of self-neglect. We were able to monitor the care recipient's routine of basic (hygiene, sleep) and instrumental (cooking, outing) activities of daily living. We showed that nonintrusive AAL monitoring can identify and present the relevant trends in daily routines in data format. This continuously gathered information can then be integrated with other sources of information to help care professionals manage risks and develop tailored intervention plans.

Replication of this study will be required to strengthen the findings of this study. Future studies also need to consider the economic benefits of accessible AAL monitoring for clinical decision making in public and private health services. Such

efforts are driven by a pressing need to deliver efficient home care services, enhance the quality of life of home care recipients, and reduce the burden on informal caregivers. We believe these results show that with further development and broader implementation, AAL monitoring systems will become essential tools in the promotion of aging in place.

Acknowledgments

This study was supported by the Centre de recherche de l'Institut universitaire de gériatrie de Montréal–Comité aviseur pour la recherche clinique and the Canadian Institutes of Health Research –Natural Sciences and Engineering Research Council. ML was supported by a postdoctoral award from the Fonds de la recherche du Québec—Santé (FRQS). NB was supported by a research scholar award from FRQS. The authors would like to thank Marie-Michèle Hachée for her valuable contribution to data collection.

Conflicts of Interest

None declared.

References

1. World Alzheimer Report 2019 Attitudes to Dementia. Alzheimer's Disease International. 2019. URL: https://www.alz.co.uk/research/WorldAlzheimerReport2019.pdf [accessed 2020-10-20] [WebCite Cache ID https://www.alz.co.uk/research/WorldAlzheimerReport2019.pdf]
2. Alzheimer's Association. 2016 Alzheimer's disease facts and figures. Alzheimers Dement 2016 Apr;12(4):459-509. [doi: 10.1016/j.jalz.2016.03.001] [Medline: 27570871]
3. Global Action Plan On the Public Health Response to Dementia 2017-2025. World Health Organization. 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/259615/9789241513487-eng.pdf [accessed 2020-10-20]
4. Marshall GA, Amariglio RE, Sperling RA, Rentz DM. Activities of daily living: where do they fit in the diagnosis of Alzheimer's disease? Neurodegener Dis Manag 2012 Oct 1;2(5):483-491 [FREE Full text] [doi: 10.2217/nmt.12.55] [Medline: 23585777]
5. Iris M, Ridings JW, Conrad KJ. The development of a conceptual model for understanding elder self-neglect. Gerontologist 2010 Jun;50(3):303-315. [doi: 10.1093/geront/gnp125] [Medline: 19726732]
6. Dong X, Gorbien M. Decision-making capacity: the core of self-neglect. J Elder Abuse Negl 2005;17(3):19-36. [doi: 10.1300/j084v17n03_02] [Medline: 16931467]
7. Dong X, Simon M, Evans D. Elder self-neglect is associated with increased risk for elder abuse in a community-dwelling population: findings from the Chicago health and aging project. J Aging Health 2013 Feb;25(1):80-96. [doi: 10.1177/0898264312467373] [Medline: 23223207]
8. Dong X, Simon MA, Evans D. Elder self-neglect and hospitalization: findings from the Chicago health and aging project. J Am Geriatr Soc 2012 Feb;60(2):202-209 [FREE Full text] [doi: 10.1111/j.1532-5415.2011.03821.x] [Medline: 22283642]
9. Dong X, Simon MA, Mosqueda L, Evans DA. The prevalence of elder self-neglect in a community-dwelling population: hoarding, hygiene, and environmental hazards. J Aging Health 2012 Apr;24(3):507-524. [doi: 10.1177/0898264311425597] [Medline: 22187089]
10. Dong X, Simon MA. Association between elder self-neglect and hospice utilization in a community population. Arch Gerontol Geriatr 2013;56(1):192-198 [FREE Full text] [doi: 10.1016/j.archger.2012.06.008] [Medline: 22770866]
11. Turner A, Hochschild A, Burnett J, Zulfiqar A, Dyer CB. High prevalence of medication non-adherence in a sample of community-dwelling older adults with adult protective services-validated self-neglect. Drugs Aging 2012 Sep;29(9):741-749. [doi: 10.1007/s40266-012-0007-2] [Medline: 23018610]
12. Papaioannou EC, Räihä I, Sirkka-Liisa K. Self-neglect of the elderly. An overview. Eur J Gen Pract 2012 Sep;18(3):187-190. [doi: 10.3109/13814788.2012.688019] [Medline: 22640528]
13. Pavlou MP, Lachs MS. Self-neglect in older adults: a primer for clinicians. J Gen Intern Med 2008 Nov;23(11):1841-1846 [FREE Full text] [doi: 10.1007/s11606-008-0717-7] [Medline: 18649111]
14. Jefferson AL, Byerly LK, Vanderhill S, Lambe S, Wong S, Ozonoff A, et al. Characterization of activities of daily living in individuals with mild cognitive impairment. Am J Geriatr Psychiatry 2008 May;16(5):375-383 [FREE Full text] [doi: 10.1097/JGP.0b013e318162f197] [Medline: 18332397]
15. Jekel K, Damian M, Wattmo C, Hausner L, Bullock R, Connelly PJ, et al. Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. Alzheimers Res Ther 2015;7(1):17 [FREE Full text] [doi: 10.1186/s13195-015-0099-0] [Medline: 25815063]
16. Marshall GA, Rentz DM, Frey MT, Locascio JJ, Johnson KA, Sperling RA, Alzheimer's Disease Neuroimaging Initiative. Executive function and instrumental activities of daily living in mild cognitive impairment and Alzheimer's disease. Alzheimers Dement 2011 May;7(3):300-308 [FREE Full text] [doi: 10.1016/j.jalz.2010.04.005] [Medline: 21575871]

17. Teng E, Becker BW, Woo E, Cummings JL, Lu PH. Subtle deficits in instrumental activities of daily living in subtypes of mild cognitive impairment. Dement Geriatr Cogn Disord 2010;30(3):189-197 [FREE Full text] [doi: 10.1159/000313540] [Medline: 20798539]
18. Luppa M, Luck T, Weyerer S, Hans-Helmut K, Brähler E, Riedel-Heller SG. Prediction of institutionalization in the elderly. A systematic review. Age Ageing 2010 Jan;39(1):31-38. [doi: 10.1093/ageing/afp202] [Medline: 19934075]
19. Wattmo C, Wallin AK, Londos E, Minthon L. Risk factors for nursing home placement in Alzheimer's disease: a longitudinal study of cognition, ADL, service utilization, and cholinesterase inhibitor treatment. Gerontologist 2011 Feb;51(1):17-27. [doi: 10.1093/geront/gnq050] [Medline: 20562471]
20. Chan M, Campo E, Estève D, Fourniols J. Smart homes: current features and future perspectives. Maturitas 2009 Oct 20;64(2):90-97. [doi: 10.1016/j.maturitas.2009.07.014] [Medline: 19729255]
21. Mattimore T, Wenger N, Desbiens N, Teno J, Hamel M, Liu H, et al. Surrogate and physician understanding of patients' preferences for living permanently in a nursing home. J Am Geriatr Soc 1997 Jul;45(7):818-824. [doi: 10.1111/j.1532-5415.1997.tb01508.x] [Medline: 9215332]
22. Nikmat AW, Al-Mashoor SH, Hashim NA. Quality of life in people with cognitive impairment: nursing homes versus home care. Int Psychogeriatr 2015 May;27(5):815-824. [doi: 10.1017/S1041610214002609] [Medline: 25497589]
23. Olsen C, Pedersen I, Bergland A, Enders-Slegers M, Jøranson N, Calogiuri G, et al. Differences in quality of life in home-dwelling persons and nursing home residents with dementia - a cross-sectional study. BMC Geriatr 2016 Jul 11;16:137 [FREE Full text] [doi: 10.1186/s12877-016-0312-4] [Medline: 27400744]
24. Bevilacqua R, Ceccacci S, Germani M, Iualè M, Mengoni M, Papetti A. Ambient assisted living. In: Longhi S, Siciliano P, Germani M, Monteriù A, editors. Smart Object for AAL: A Review. New York, USA: Springer; 2014:313-324.
25. Kang W, Shin D, Shin D. Detecting and Predicting of Abnormal Behavior Using Hierarchical Markov Model in Smart Home Network. In: 2010 IEEE 17Th International Conference on Industrial Engineering and Engineering Management. 2010 Presented at: IEEE'10; October 29-31, 2010; Xiamen, China. [doi: 10.1109/icieem.2010.5646583]
26. Peetoom KK, Lexis MA, Joore M, Dirksen CD, de Witte LP. Literature review on monitoring technologies and their outcomes in independently living elderly people. Disabil Rehabil Assist Technol 2015 Jul;10(4):271-294. [doi: 10.3109/17483107.2014.961179] [Medline: 25252024]
27. Lussier M, Couture M, Moreau M, Laliberté C, Giroux S, Pigot H, et al. Integrating an ambient assisted living monitoring system into clinical decision-making in home care: an embedded case study. Gerontechnology 2020 Mar 1;19(1):77-92 [FREE Full text] [doi: 10.4017/gt.2020.19.1.008.00]
28. Kaye J. Home-based technologies: a new paradigm for conducting dementia prevention trials. Alzheimers Dement 2008 Jan;4(1 Suppl 1):S60-S66. [doi: 10.1016/j.jalz.2007.10.003] [Medline: 18632003]
29. Al-Shaqi R, Mourshed M, Rezgui Y. Progress in ambient assisted systems for independent living by the elderly. Springerplus 2016;5:624 [FREE Full text] [doi: 10.1186/s40064-016-2272-8] [Medline: 27330890]
30. Lussier M, Lavoie M, Giroux S, Consel C, Guay M, Macoir J, et al. Early detection of mild cognitive impairment with in-home monitoring sensor technologies using functional measures: a systematic review. IEEE J Biomed Health Inform 2019 Mar;23(2):838-847. [doi: 10.1109/jbhi.2018.2834317]
31. Parsons TD, Kane RL. Computational Neuropsychology: Current and Future Prospects for Interfacing Neuropsychology and Technology. APA PsycNet Advanced Search - PsycNET. 2017. URL: https://psycnet.apa.org/record/2017-28107-016 [accessed 2020-10-22]
32. Chikhaoui B, Lussier M, Gagnon M, Pigot H, Giroux S, Bier N. Automatic Identification of Behavior Patterns in Mild Cognitive Impairments and Alzheimer's Disease Based on Activities of Daily Living. In: International Conference on Smart Homes and Health Telematics. 2017 Presented at: ICOST'18; February 15, 2018; Singapore. [doi: 10.1007/978-3-319-94523-1_6]
33. Dawadi PN, Cook DJ, Schmitter-Edgecombe M, Parsey C. Automated assessment of cognitive health using smart home technologies. Technol Health Care 2013;21(4):323-343 [FREE Full text] [doi: 10.3233/THC-130734] [Medline: 23949177]
34. Dawadi PN, Cook DJ, Schmitter-Edgecombe M. Automated cognitive health assessment using smart home monitoring of complex tasks. IEEE Trans Syst Man Cybern Syst 2013 Nov;43(6):1302-1313 [FREE Full text] [doi: 10.1109/TSMC.2013.2252338] [Medline: 25530925]
35. Fredericks E, Bowers K, Price K, Hariri R. CAL: A Smart Home Environment for Monitoring Cognitive Decline. In: 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). 2018 Presented at: ICDCS'18; July 2-5, 2018; Vienna, Austria. [doi: 10.1109/ICDCS.2018.00155]
36. Paudel R, Dunn K, Eberle W, Chaung D. Cognitive Health Prediction on the Elderly Using Sensor Data in Smart Homes. In: The Thirty-First International Flairs Conference. 2018 Presented at: AAAI'18; May 21-23, 2018; Melbourne.
37. Card D. The Black Box Metaphor in Machine Learning. Towards Data Science. 2017. URL: https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0 [accessed 2020-10-07]
38. Castelvecchi D. Can we open the black box of AI? Nature 2016 Oct 6;538(7623):20-23. [doi: 10.1038/538020a] [Medline: 27708329]
39. Kazdin A. Single-Case Research Designs: Methods for Clinical and Applied Settings. New York, USA: Oxford University Press; 2011.

40. Ordre des infirmières et infirmiers du Québec. Infirmière Clinicienne Ou Infirmier Clinicien. OIIQ: Ordre Des Infirmières Et Infirmiers Du Québec. 2020. URL: https://www.oiiq.org/acceder-profession/decouvrir-la-profession/possibilites-de-carriere/infirmiere-clinicienne-ou-infirmier-clinicien [accessed 2020-10-07]
41. Rantz M, Skubic M, Koopman R, Phillips L, Alexander G, Miller S. Using Sensor Networks to Detect Urinary Tract Infections in Older Adults. In: 13th International Conference on e-Health Networking, Applications and Services. 2011 Presented at: IEEE'11; June 13-15, 2011; Columbia, USA. [doi: 10.1109/health.2011.6026731]
42. da Silva RA. Sleep disturbances and mild cognitive impairment: a review. Sleep Sci 2015;8(1):36-41 [FREE Full text] [doi: 10.1016/j.slsci.2015.02.001] [Medline: 26483941]
43. Akl A, Chikhaoui B, Mattek N, Kaye J, Austin D, Mihailidis A. Clustering home activity distributions for automatic detection of mild cognitive impairment in older adults. J Ambient Intell Smart Environ 2016;8(4):437-451 [FREE Full text] [doi: 10.3233/AIS-160385] [Medline: 27617044]
44. Dawadi PN, Cook DJ, Schmitter-Edgecombe M. Modeling patterns of activities using activity curves. Pervasive Mob Comput 2016 Jun;28:51-68 [FREE Full text] [doi: 10.1016/j.pmcj.2015.09.007] [Medline: 27346990]
45. Suzuki T, Murase S, Tanaka T, Okazawa T. New approach for the early detection of dementia by recording in-house activities. Telemed J E Health 2007 Feb;13(1):41-44. [doi: 10.1089/tmj.2006.0033] [Medline: 17309353]
46. Ju YS, McLeland JS, Toedebusch CD, Xiong C, Fagan AM, Duntley SP, et al. Sleep quality and preclinical Alzheimer disease. JAMA Neurol 2013 May;70(5):587-593 [FREE Full text] [doi: 10.1001/jamaneurol.2013.2334] [Medline: 23479184]
47. Farrell MT, Zahodne LB, Stern Y, Dorrejo J, Yeung P, Cosentino S. Subjective word-finding difficulty reduces engagement in social leisure activities in Alzheimer's disease. J Am Geriatr Soc 2014 Jun;62(6):1056-1063 [FREE Full text] [doi: 10.1111/jgs.12850] [Medline: 24890186]
48. Ghisletta P, Bickel J, Lövdén M. Does activity engagement protect against cognitive decline in old age? methodological and analytical considerations. J Gerontol B Psychol Sci Soc Sci 2006 Sep;61(5):P253-P261. [doi: 10.1093/geronb/61.5.p253] [Medline: 16960228]
49. Hughes TF, Flatt JD, Fu B, Chang CH, Ganguli M. Engagement in social activities and progression from mild to severe cognitive impairment: the MYHAT study. Int Psychogeriatr 2013 Apr;25(4):587-595 [FREE Full text] [doi: 10.1017/S1041610212002086] [Medline: 23257280]
50. Small BJ, Dixon RA, McArdle JJ, Grimm KJ. Do changes in lifestyle engagement moderate cognitive decline in normal aging? Evidence from the Victoria longitudinal study. Neuropsychology 2012 Mar;26(2):144-155 [FREE Full text] [doi: 10.1037/a0026579] [Medline: 22149165]
51. Senturk G, Bilgic B, Arslan AB, Bayram A, Hanagasi H, Gurvit H, et al. Cognitive and anatomical correlates of anosognosia in amnestic mild cognitive impairment and early-stage Alzheimer's disease. Int Psychogeriatr 2017 Feb;29(2):293-302. [doi: 10.1017/S1041610216001812] [Medline: 27780496]
52. Sartorius A, Aksay SS, Hausner L, Frölich L. Severe agitation in severe early-onset Alzheimer’s disease resolves with ECT. Neuropsychiatr Dis Treat 2014 Nov:2147. [doi: 10.2147/ndt.s71008]
53. Wesson J, Luchins J. Empirical evaluation of the global deterioration scale for staging Alzheimer's disease. J Am J Psychiatry 1993 Apr;150(4):680-682. [doi: 10.1176/ajp.150.4.680]
54. Boscart VM, McNeill S, Grinspun D. Dementia care in Canada: nursing recommendations. Can J Aging 2019 Sep;38(3):407-418. [doi: 10.1017/S071498081800065X] [Medline: 31385569]
55. Mahoney EL, Mahoney DF. Acceptance of wearable technology by people with Alzheimer's disease: issues and accommodations. Am J Alzheimers Dis Other Demen 2010 Sep;25(6):527-531. [doi: 10.1177/1533317510376944] [Medline: 20702501]
56. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313. [doi: 10.1177/1357633X20916567] [Medline: 32196391]

Abbreviations

AAL: ambient assisted living
AD: Alzheimer disease
FRQS: Fonds de la recherche du Québec—Santé
PIR: passive infrared

Edited by C Lovis; submitted 09.07.20; peer-reviewed by J Li, N Mohammad Gholi Mezerji; comments to author 28.07.20; revised version received 10.08.20; accepted 26.09.20; published 13.11.20.

Please cite as:
Lussier M, Aboujaoudé A, Couture M, Moreau M, Laliberté C, Giroux S, Pigot H, Gaboury S, Bouchard K, Belchior P, Bottari C, Paré G, Consel C, Bier N
Using Ambient Assisted Living to Monitor Older Adults With Alzheimer Disease: Single-Case Study to Validate the Monitoring Report
JMIR Med Inform 2020;8(11):e20215
URL: https://medinform.jmir.org/2020/11/e20215
doi:10.2196/20215
PMID:33185555

Original Paper

Visualization Environment for Federated Knowledge Graphs: Development of an Interactive Biomedical Query Language and Web Application Interface

Steven Cox[1], BS; Stanley C Ahalt[1], PhD; James Balhoff[1], PhD; Chris Bizon[1], PhD; Karamarie Fecho[1], PhD; Yaphet Kebede[1], MS; Kenneth Morton[2], PhD; Alexander Tropsha[1,3], PhD; Patrick Wang[2], PhD; Hao Xu[1], PhD

[1]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]CoVar Applied Technologies, Durham, NC, United States

[3]UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Corresponding Author:
Steven Cox, BS
Renaissance Computing Institute
University of North Carolina at Chapel Hill
100 Europa Drive
Suite 540
Chapel Hill, NC, 27517
United States
Phone: 1 (919) 445 9640
Email: scox@renci.org

Abstract

Background: Efforts are underway to semantically integrate large biomedical knowledge graphs using common upper-level ontologies to federate graph-oriented application programming interfaces (APIs) to the data. However, federation poses several challenges, including query routing to appropriate knowledge sources, generation and evaluation of answer subsets, semantic merger of those answer subsets, and visualization and exploration of results.

Objective: We aimed to develop an interactive environment for query, visualization, and deep exploration of federated knowledge graphs.

Methods: We developed a biomedical query language and web application interphase—termed as Translator Query Language (TranQL)—to query semantically federated knowledge graphs and explore query results. TranQL uses the Biolink data model as an upper-level biomedical ontology and an API standard that has been adopted by the Biomedical Data Translator Consortium to specify a protocol for expressing a query as a graph of Biolink data elements compiled from statements in the TranQL query language. Queries are mapped to federated knowledge sources, and answers are merged into a knowledge graph, with mappings between the knowledge graph and specific elements of the query. The TranQL interactive web application includes a user interface to support user exploration of the federated knowledge graph.

Results: We developed 2 real-world use cases to validate TranQL and address biomedical questions of relevance to translational science. The use cases posed questions that traversed 2 federated Translator API endpoints: Integrated Clinical and Environmental Exposures Service (ICEES) and Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways (ROBOKOP). ICEES provides open access to observational clinical and environmental data, and ROBOKOP provides access to linked biomedical entities, such as “gene,” “chemical substance,” and “disease,” that are derived largely from curated public data sources. We successfully posed queries to TranQL that traversed these endpoints and retrieved answers that we visualized and evaluated.

Conclusions: TranQL can be used to ask questions of relevance to translational science, rapidly obtain answers that require assertions from a federation of knowledge sources, and provide valuable insights for translational research and clinical practice.

(JMIR Med Inform 2020;8(11):e17964) doi:10.2196/17964

KEYWORDS

knowledge graphs; clinical data; biomedical data; federation; ontologies; semantic harmonization; visualization; application programming interface; translational science; clinical practice

Introduction

The semantic web is comprised of a collection of large graph databases of knowledge assertions enumerated as subject-predicate-object "triples" [1]. The subject and the object are considered unique entities that have globally unique identifiers. The predicate provides a relationship between the subject and the object. These assertions typically are derived from costly, manual digital curation of data sources in order to record both the assertion and the provenance of each assertion, although tools such as DataStaR have been developed to support efficiency and scale [2].

Recent work aims to semantically federate these "knowledge graphs" (KGs) in a manner that supports a unified query interface to address the heterogeneity of the corpus of knowledge sources [3]. These efforts have employed varied technological tools, including open web application programming interfaces (APIs), graph databases, and interface standards. The federated design of core semantic web technologies (such as SPARQL) has long acknowledged that the domain of all interconnected knowledge is too large to manage via a single monolithic database infrastructure with a single curatorial staff. Rather, the ability to allow multiple teams to work independently, while connecting them through semantic consensus, supports scale and the independence to experiment with new kinds of data not envisioned by a monolithic system. Federation provides such an environment to support a diverse community of collaborators, foster modular disciplinary specialization, and integrate multiple knowledge sources, both public and private.

However, federation presents challenges, in that any query system must provide semantic query planning to route queries to the appropriate knowledge sources, generate and evaluate answer subsets, and then merge the answer subsets into a semantically valid, coherent whole.

The Biomedical Data Translator program (hereinafter referred to as "Translator") [4-6], funded by the National Center for Advancing Translational Sciences, National Institutes of Health, was created to address the challenge of query and interrogation across diverse data types and the many open data sets that are available but are not semantically compatible. The ability to interrogate relationships across the full spectrum of data types is critical in order to find answers to pressing translational questions. The Translator semantic informatics platform provides a comprehensive, unified semantic framework and approach to support such interrogation across disparate knowledge sources.

Herein, we present the Translator Query Language (TranQL) as a graph-oriented biomedical query language and web application to support query and visualization of semantically federated biomedical knowledge sources for deep, iterative exploration of query answers. TranQL leverages the semantic framework developed as part of the Translator program to support query across the Translator ecosystem.

Methods

Overview of TranQL Design

TranQL was designed to overcome the challenges in federating queries across heterogeneous distributed knowledge sources and merging answer subsets into a semantically cohesive whole. A key design feature is the adoption of a shared schema or namespace for navigating a globally agreed-upon conceptual structure that expresses entity types and the relationships between them. Such a schema accommodates queries that target subsets of the larger federated space and provides flexibility to extend domains or specialize within a specific domain. TranQL leverages the Biolink data model and the Translator Knowledge Graph Standard (KGS) API to implement this design feature. The TranQL query language and the TranQL web application are used for query execution and exploration of query results. These components are described in detail in the following sections.

Biolink Data Model

The Biolink data model [7] provides the highest level of abstraction of biomedical concepts, thereby omitting the level of specificity elaborated by ontologies specific to biomedical domains such as those focused solely on "disease," "phenotype," or "anatomy." Biolink is hierarchical and addresses both entities and relationships between them. TranQL uses the Biolink data model as an upper-level biomedical ontology to express concepts and relationships in the body of a knowledge query.

Translator KGS API

The Translator KGS API [8] was developed as part of the Translator program and specifies a standard protocol for expressing a query as a graph of Biolink data elements that can be resolved into a KG consisting of nodes and edges from multiple knowledge sources and mappings between the KG and specific elements of the query. The Translator KGS API, therefore, enables a semantically coherent and technologically uniform query pattern over a federated, but otherwise heterogeneous, knowledge network.

TranQL Query Language

The TranQL query language consists of a lexical analyzer, parser, and abstract syntax tree. The lexical analyzer recognizes a query language with the structure shown in Figure 1.

Figure 1. Structure of the query language.

```
<SET> ::= <variable> = <value>
<SELECT> ::= SELECT <graph> FROM <service> [WHERE <constraint> [AND <constraint]*]
             [[SET <jsonpath> AS <var> | [SET <var>]]*
<CREATE-GRAPH> ::= <var> AT <service> AS <name>
```

The "set" command assigns a value to a variable. The "select" command specifies a query graph of linked Biolink concepts. The query graph provides the framework for the resultant federated knowledge network. The "from" clause specifies the Translator KGS API. The invoked Translator KGS API endpoints return nodes and edges that conform to the semantic structure defined by the query graph. The optional "where" clause begins a list of constraints. Constraints can specify a value for an element of the query graph, supply options for the service invocation, or filter results. The "select" command may include a "set" command. For example, one form uses a JSON path query to extract an element of the resultant KG and assign it to a variable; another form assigns the entire KG to a variable. Finally, the "create graph" statement sends the resultant named KG to a service for storage.

TranQL Web Application

The TranQL web application includes several components that together support interactive exploration: the TranQL query schema, TranQL backplane, and TranQL user interface (UI). The TranQL UI is a single-page web application that includes a query editor, cache function, graph visualization environment, answer viewer, and visualization controls.

TranQL Schema

The TranQL schema declares Translator KGS API endpoints and associates each one with a graph of the kinds of entity transitions it supports. These semantic transition "maps" are obtained from each endpoint. They describe the capabilities of endpoints in terms of the Biolink data model. Each endpoint is represented as a hierarchy in which the first level is an entity, the second is an entity, and the third is a relationship between the 2 entities. TranQL's schema merges these transition maps into a unified KG. "Select" statements, containing the schema endpoint in the "from" clause, enable query planning. Planning compares each step in the query graph with the schema's possible transitions to construct a plan for service invocations. The engine executes the plan by sending fragments of the query graph to those services able to complete the request. It then passes the results from one query segment to the next query segment and merges the composite results. Importantly, metadata indicating the provenance of each node and edge in the graph are preserved.

TranQL Backplane

The TranQL backplane is an OpenAPI implemented as a protocol normalization layer over the federated Translator KGS API endpoints. The backplane standardizes incompatibilities and separates the details of service invocations from the query language. When invoking backplane services, the "from" clause specifies an abbreviated syntax, thereby avoiding the need to specify the HTTP protocol, domain name, and port syntax to a specific service. The TranQL OpenAPI processes TranQL queries by parsing the query, planning service-specific questions that are implied by the query, executing those queries by invoking the backplane, collecting answers from services, and merging those answers into a single Translator KGS API–compliant response.

Query Editor and Cache Function

The query editor provides syntax highlighting, line numbers, and keyword autocompletion for the TranQL query language and the Biolink data model. The editor is implemented using CodeMirror [9], which provides the basis for many TranQL capabilities. Running a query retrieves a cached answer from a previous execution of the query or invokes one or more Translator KGS API endpoints. The resultant KG is rendered in a force-directed graph layout. Results are cached using the text of the query as a key. Changes to the query circumvent the cache. The query cache can be temporarily disabled or completely cleared via the settings dialog.

Graph Visualization Environment

The graph visualization environment uses the graphical processing unit (GPU)–accelerated 3D-rendering components based on WebGL [10] and Three.js [11]. These tools support fluid exploration of KGs comprised of tens of thousands of nodes. TranQL uses a configurable force-directed layout to control the distance between nodes, thereby providing a more intuitive visual organization of the KG. Nodes and edges are colored according to semantic type, as defined by the Biolink data model. Mouse events trigger the display of available information on each component of the KG. Deeper investigation of each node is supported by the "Select" mode. In "Navigation" mode, the selection of any node in the KG will center the selected node and orient the camera to look directly at it, thus obviating the need for incremental manual navigation.

Answer Viewer

The TranQL application integrates the ROBOKOP (Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways) answer viewer [12] to render a tabular view of the KG, as opposed to an interactive graph view. Both TranQL and ROBOKOP adhere to the Translator KGS API protocol, thereby allowing TranQL to send query results directly to the ROBOKOP answer viewer.

Visualization Controls

The "Settings" dialog provides 3 control panels for visualizations. The first panel allows users to select any of the 2-dimensional (2D), 3D, or virtual reality as the display style for the KG, with toggling to choose whether nodes and edges are colored by type. This panel also has controls for temporarily disabling the cache or clearing the cache entirely. A second panel provides 2 controls that relate to the structure of the KG. The first control configures a range of edge weights such that only edges with weights falling within the range will be rendered in the visualization. The second control configures a range of node connectivity such that only nodes with connections falling within the range will be rendered in the visualization. The third panel presents a list of checkboxes that correspond to knowledge sources that provide edges in the answer KG. The checkboxes allow users to select the preferred knowledge sources for rendering in the visualization.

Results

Overview of Use Case Results

We developed use cases in which a single query was posed to the TranQL UI and spanned federated knowledge managed by 2 independent Translator KGS API endpoints. The first endpoint was ICEES (Integrated Clinical and Environmental Exposures Service [13,14]. ICEES provides open access to clinical data derived from UNC Health that have been integrated with public data on environmental exposures. The second endpoint was ROBOKOP [12,15,16]. ROBOKOP is an open question–answering system that provides access to linked biomedical entities, such as "gene," "chemical substance," and "disease," that are derived largely from curated public data sources.

Use Case 1

The first query asked in natural language, "What diseases are differentially associated with males and females, and what genes and chemical substances are associated with those diseases?" The overall intent of the query was to (1) validate TranQL by replicating established findings on sex differences and (2) discover chemicals or drugs that may be used to treat diseases that differentially affect males versus females. The natural language question is manually translated into the TranQL query language, and the resultant query is shown in Figure 2.

Figure 2. TranQL query for use case 1.

```
SELECT population_of_individual_organisms->disease->gene->chemical_substance
FROM "/schema"
WHERE icees.table = 'patient'
AND icees.year = 2010
AND icees.feature.Sex2 = 'Female'
AND icees.maximum_p_value = 0.5
```

The first part of the query derives validation from ICEES on diseases differentially diagnosed in males versus females (ie, greater than chance); the second part of the question derives exploratory information from ROBOKOP on genes and chemical substances associated with those diseases. The "from" clause directs TranQL to conduct schema planning. The "icees" prefixed parameters in the "where" clause are feature variables within ICEES and become options in the Translator KGS API protocol that are transmitted to the schema element designated by each parameter.

The TranQL query, as posed to the UI, and the resultant answer KG are shown in Figure 3. Note that nodes are color-coded according to entity type; links are similarly color-coded. Users can explore the answer KG using a variety of tools to enable zooming, rotation, choice of KG views (2D, 3D, and virtual reality), a tabular view of connected nodes, and a variety of other features.

One pathway in the answer KG shows that males and females are differentially diagnosed with ovarian cancer, a known female disease, thus validating the ability to extract factual information from ICEES using TranQL. The pathway traverses from ICEES to ROBOKOP to demonstrate an association between ovarian cancer and the gene *PTH* (parathyroid hormone), which ROBOKOP associates with the chemical substances—calcitriol, calcium atom, vitamin D, calcium carbonate, phosphane, adenine, phosphate, phosphorous, maxacalcitol, calciol, calcium, lithium hydride, and cinacalcet. Figure 4 highlights *PTH* and demonstrates the "object viewer" feature that allows users to review the metadata associated with the gene.

A quick Google search identifies several case reports describing hypercalcemia associated with parathyroid hormone in women with ovarian cancer, including a high-profile study by Nussbaum and colleagues [17] and a more recent one by Ma and colleagues [18]. These findings provide validation of TranQL and further suggest avenues for exploratory drug discovery in the treatment of ovarian cancer.

Figure 3. Use Case 1: example TranQL query and answer KG. The TranQL query was manually translated into the TranQL query language from a natural language query that asked, "What diseases are differentially associated with males and females, and what genes and chemical substances are associated with those diseases?".

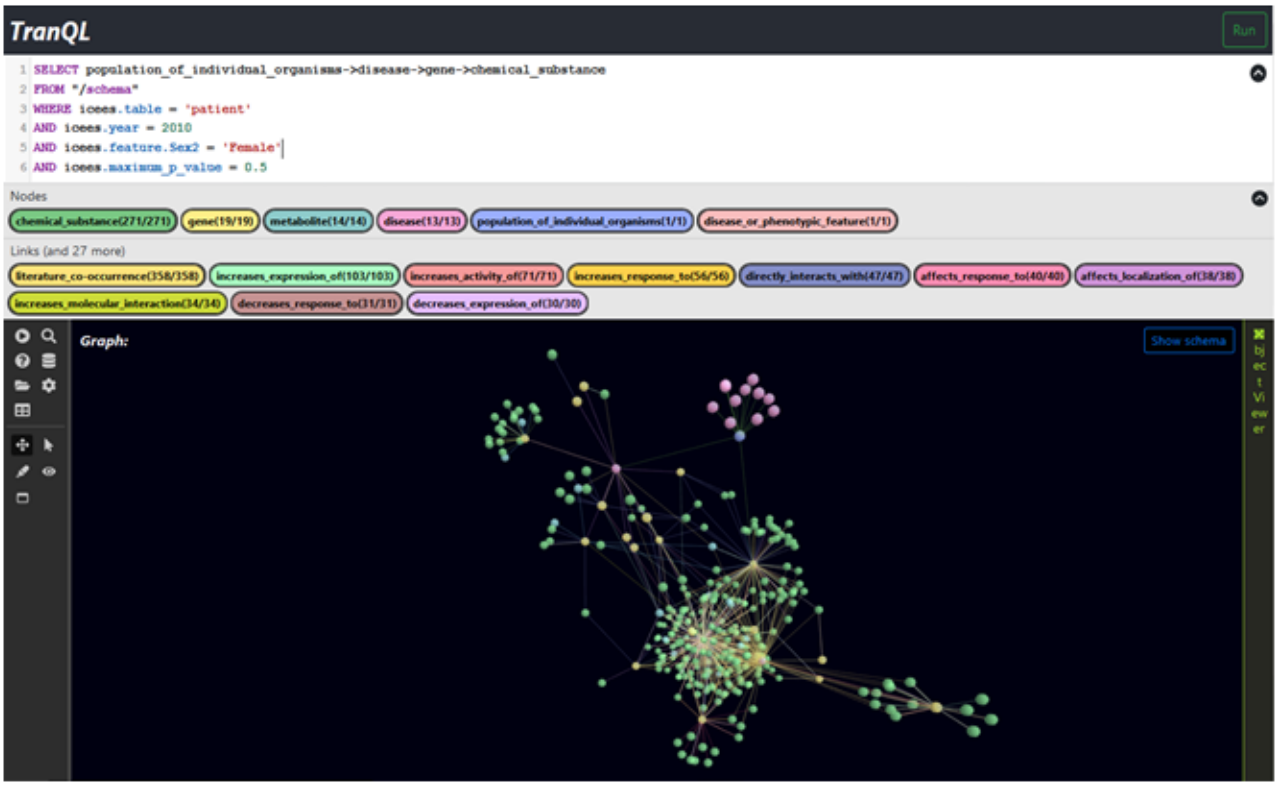
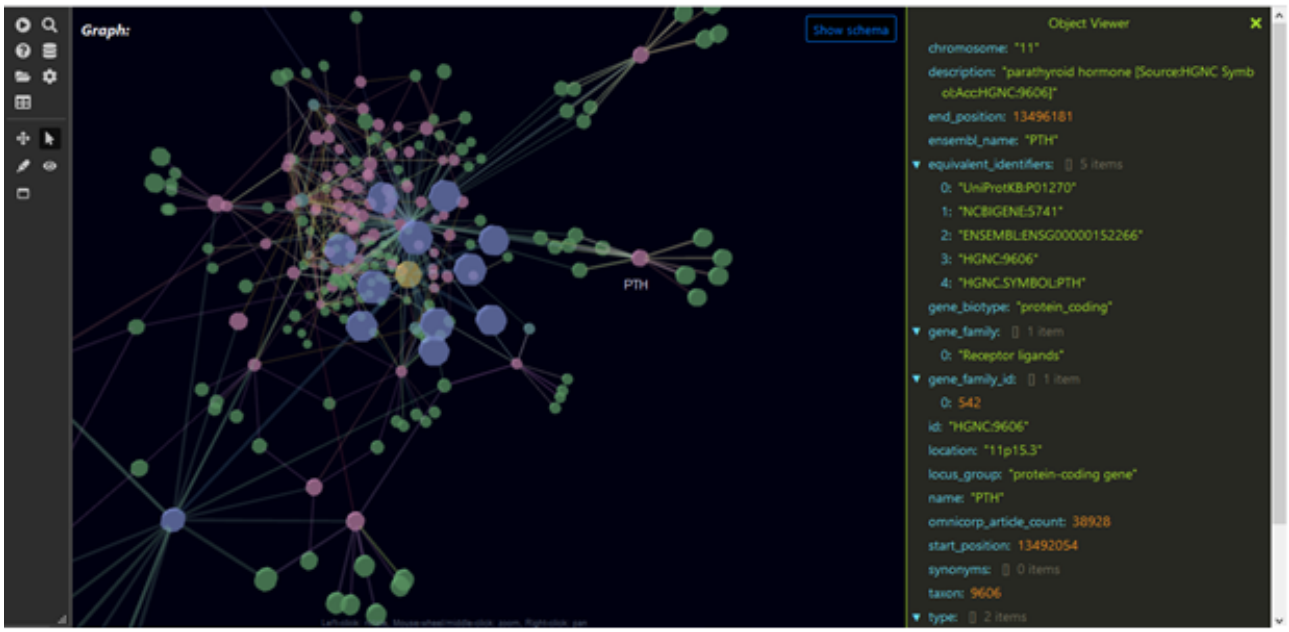


Figure 4. Metadata associated with the gene *PTH* (parathyroid hormone), which are found in the answer KG for the TranQL query shown in Figure 3.



Use Case 2

A second example query asked in natural language, "What diseases are differentially expressed in patients who live in rural versus urban regions, and what genes and chemical substances are associated with those diseases?" As with the first use case, the natural language query in this use case is manually translated into the TranQL query language, as shown in Figure 5.

As with the first example query, this query also extracts clinical information on diseases from ICEES and biomedical information on genes and chemical substances from ROBOKOP. The query, as posed in the UI, and the resultant answer KG are shown in Figure 6.

Figure 5. TranQL query for use case 2.

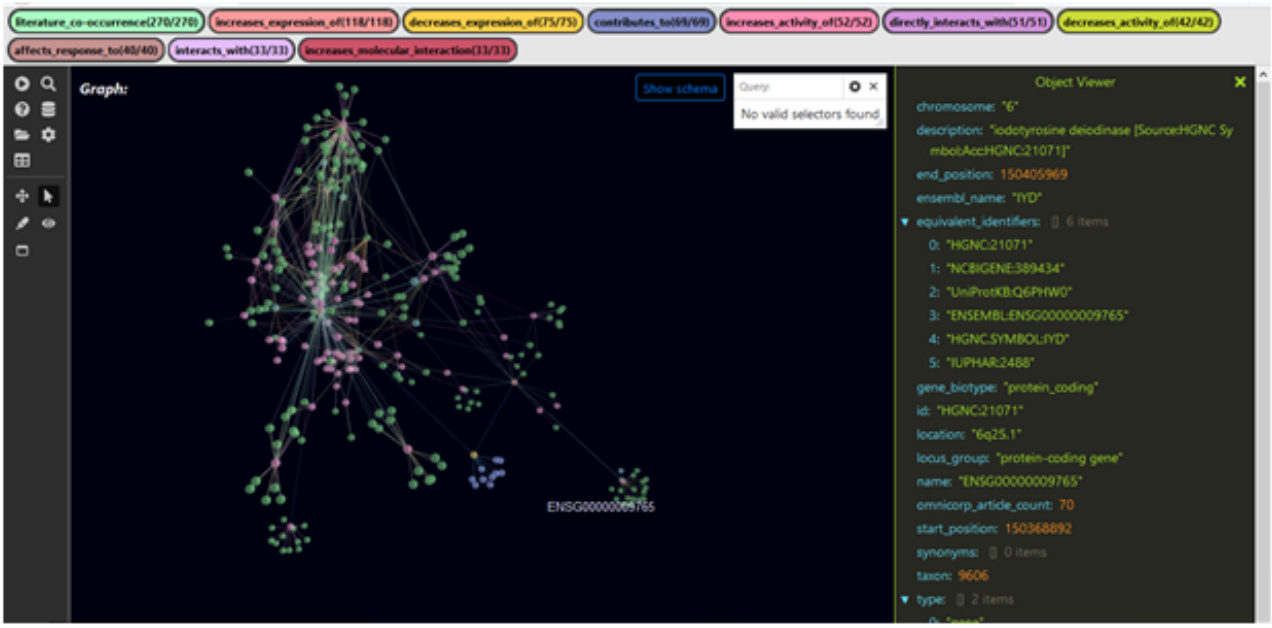
```
SELECT population_of_individual_organisms->disease->gene->chemical_substance
FROM "/schema"
WHERE icees.table = 'patient'
AND icees.year = 2010
AND icees.feature.EstResidentialDensity = 1
AND icees.maximum_p_value = 0.5
```

Figure 6. Use Case 2: example TranQL query and resultant KG. The TranQL query was manually translated into the TranQL query language from a natural language query that asked, "What diseases are differentially expressed in patients who live in rural versus urban regions, and what genes and chemical substances are associated with those diseases?".

One pathway through the resultant answer KG shows that croup is differentially diagnosed among patients who live in rural versus urban regions. The pathway traverses from ICEES to ROBOKOP and identifies the gene *IYD* (iodotyrosine deiodinase), as shown in Figure 7. ROBKOP associates *IYD* with the chemicals—polybrominated biphenyls, pentabromodiphenyl ether, halogenated diphenyl ethers, 2,2',4,5'-tetrabromodiphenyl ether, 3,5-diiodo-L-tyrosine, triclosan, trihydroiodine, erythrosine, rose bengal, hydrogen iodide, benzbromarone, chlorobiphenyl, fenson, NADPH, NADP(+), flavin mononucleotide, and eosin B.

Several quick Google searches find that the majority of the identified chemicals represent hazardous substances (eg, polybrominated biphenyls) or food additives (eg, flavin mononucleotide) that differentially affect infants and young children, either due to enhanced toxicity or increased probability of exposure. For instance, polybrominated biphenyls are classified by the US Environmental Protection Agency as hazardous substances that were once used as flame retardants and plastic additives [19]. Although they are now prohibited, the compounds remain in the environment and differentially impact infants and young children. Similarly, croup differentially affects infants and young children [20], thus providing a plausible explanation for the association and further suggesting that the identified compounds may also contribute directly to croup.

Figure 7. Metadata associated with the gene *IYD* (iodotyrosine deiodinase), which are found in the answer KG for the TranQL query shown in Figure 6.



Discussion

Principal Findings

We have created a biomedical query language—TranQL—to support query, visualization, and exploration of federated biomedical KGs. TranQL leverages the semantic framework and approach developed as part of the Translator program to semantically integrate large biomedical KGs, thereby allowing users to pose challenging translational queries that span multiple knowledge sources, rapidly explore the results, and derive valuable insights for translational research. Importantly, we have validated TranQL using 2 driving use cases and queries that span clinical knowledge sources and observational biomedical knowledge sources.

The Translator semantic platform allows users to interrogate relationships across the full spectrum of data types, without needing to manually search through individual databases and data sets that exhibit varying levels of semantic inference rules and linkages among entities. Moreover, the Translator framework supports consistent data linkage and semantic resolution across data sources due to adoption of the Biolink data model and the Translator KGS API specification as well as mappings to relevant biomedical ontologies, such as Monarch Disease Ontology and Human Phenotype Ontology.

In addition to leveraging the Translator framework and approach, TranQL provides several other capabilities. First, TranQL offers a domain-specific query language that makes iterative query and result exploration practical and approachable by users without any experience of software development. TranQL also provides a layer of abstraction for accessing a federation of services, and the interactive web interface offers graphical and tabular visualizations of query answers. Moreover, although not demonstrated herein, TranQL is designed to scale to federations involving more than 2 knowledge sources.

Limitations

TranQL has several limitations that should be considered. First, TranQL is constrained in expressivity, in that the tool can only represent linear paths through KGs, as opposed to more complex structures. Second, TranQL is limited in the number of available features and the ability to perform operations. Third, users are required to manually map natural language queries to the TranQL query language. Automated machine translation of natural language queries to machine queries remains a major challenge within the Translator program and elsewhere, primarily due to ambiguity related to intent and context [21]. Fourth, unstructured data likewise remain a major challenge within the Translator program and elsewhere, although progress has been made in certain areas. For instance, Translator team members are developing tools to handle notoriously challenging types of unstructured data such as clinical laboratory measures, including the LOINC2HPO tool that maps clinical laboratory measures to the human phenotype ontology [22]. Fifth, TranQL answer sets can be challenging to navigate, especially when a large number of nodes and edges are returned to users. This remains a challenge with KGs in general, although we are considering approaches to provide additional views such as the rudimentary tabular view that is under development. Finally, although KGs are a common approach to knowledge representation, they are by no means the only approach. For instance, knowledge fusion patterns are gaining in popularity as a new form of knowledge representation [23].

Despite these limitations, we believe that TranQL will find broad adoption due to its ability to support speed to discovery of insights to complex translational questions and generate mechanistic hypotheses for subsequent investigation and testing.

Future Directions

TranQL is under active development, with performance and feature enhancements deployed regularly. Planned feature

enhancements include a more user-friendly interface and additional visualization capabilities to support answer exploration. We are working with subject matter experts to iteratively improve the TranQL UI and other features of the interactive web application. We are also working to improve the robustness of query answers. For instance, we are developing approaches to harmonize across entity identifiers in order to accommodate the disparate identifier systems adopted by different Translator KGS API endpoints and enable more complete federation of those knowledge sources.

Availability

TranQL is publicly available on its website [24]. Software code and instructions are available under the MIT open software license at a GitHub repository [25]. We encourage users to post issues to the TranQL GitHub repository and report identified software bugs or request new features and capabilities. Additional information and example queries can be found on the webpages of TranQL API [26] and the Translator program [27].

Acknowledgments

This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (grant numbers OT3TR002020, OT2TR002514). The authors thank Matt Brush and Chris Mungall for creating and publishing the Biolink data model, and Eric Deutch and David Koslicki for contributing to the Translator KGS API specification. The authors also thank Marian Mersmann for proofreading the final version of the manuscript that was submitted for review.

Conflicts of Interest

None declared.

References

1. Rueda CA. Semantic Web: Core Concepts and Mechanisms, MMI ORR-Ontology Registry and Repository. 2019 Jan 01. URL: https://speakerdeck.com/carueda/semantic-web-core-concepts-and-mechanisms-and-mmi-orr-ontology-registry-and-repository [accessed 2020-01-23]
2. Khan H, Caruso B, Corson-Rikert J, Dietrich D, Lowe B, Steinhart G. DataStaR: Using the Semantic Web approach for Data Curation. IJDC 2011 Jul 25;6(2):209-221 [FREE Full text] [doi: 10.2218/ijdc.v6i2.197]
3. Sima AC, Mendes de Farias T, Zbinden E, Anisimova M, Gil M, Stockinger H, et al. Enabling semantic queries across federated bioinformatics databases. Database (Oxford) 2019 Jan 01;2019 [FREE Full text] [doi: 10.1093/database/baz106] [Medline: 31697362]
4. Austin CP, Colvis CM, Southall NT. Deconstructing the Translational Tower of Babel. Clin Transl Sci 2019 Mar;12(2):85 [FREE Full text] [doi: 10.1111/cts.12595] [Medline: 30412342]
5. Biomedical Data Translator Consortium. The Biomedical Data Translator Program: Conception, Culture, and Community. Clin Transl Sci 2019 Mar;12(2):91-94 [FREE Full text] [doi: 10.1111/cts.12592] [Medline: 30412340]
6. Biomedical Data Translator Consortium. Toward A Universal Biomedical Data Translator. Clin Transl Sci 2019 Mar;12(2):86-90 [FREE Full text] [doi: 10.1111/cts.12591] [Medline: 30412337]
7. Schema and generated objects for biolink data model and upper ontology. Biolink GitHub repository. URL: https://biolink.github.io/biolink-model/ [accessed 2020-01-23]
8. Translator KGS API GitHub repository. Translator KGS API Specification. URL: https://github.com/NCATS-Tangerine/NCATS-ReasonerStdAPI [accessed 2020-01-23]
9. CodeMirror. URL: https://codemirror.net/ [accessed 2020-01-23]
10. WebGL. URL: https://get.webgl.org/ [accessed 2020-01-23]
11. three.js. URL: https://threejs.org/ [accessed 2020-01-23]
12. Morton K, Wang P, Bizon C, Cox S, Balhoff J, Kebede Y, et al. ROBOKOP: an abstraction layer and user interface for knowledge graphs to support question answering. Bioinformatics 2019 Dec 15;35(24):5382-5384. [doi: 10.1093/bioinformatics/btz604] [Medline: 31410449]
13. ICEES API. URL: https://icees.renci.org:16340/apidocs/ [accessed 2020-01-23]
14. Fecho K, Pfaff E, Xu H, Champion J, Cox S, Stillwell L, et al. A novel approach for exposing and sharing clinical data: the Translator Integrated Clinical and Environmental Exposures Service. J Am Med Inform Assoc 2019 Oct 01;26(10):1064-1073 [FREE Full text] [doi: 10.1093/jamia/ocz042] [Medline: 31077269]
15. Robokop. URL: https://robokop.renci.org/ [accessed 2020-01-23]
16. Bizon C, Cox S, Balhoff J, Kebede Y, Wang P, Morton K, et al. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. J Chem Inf Model 2019 Dec 23;59(12):4968-4973. [doi: 10.1021/acs.jcim.9b00683] [Medline: 31769676]
17. Nussbaum SR, Gaz RD, Arnold A. Hypercalcemia and ectopic secretion of parathyroid hormone by an ovarian carcinoma with rearrangement of the gene for parathyroid hormone. N Engl J Med 1990 Nov 08;323(19):1324-1328. [doi: 10.1056/NEJM199011083231907] [Medline: 2215618]

18. Ma X, Wang Y, Zhang X, Dong M, Yang W, Xue F. Ovarian cancer presenting with hypercalcemia: two cases with similar manifestations but different mechanisms. Cancer Biol Med 2018 May;15(2):182-187 [FREE Full text] [doi: 10.20892/j.issn.2095-3941.2018.0009] [Medline: 29951343]
19. United States Environmental Protection Agency. Polybrominated byphenyls (PBBs). Technical Fact Sheet. URL: https://www.epa.gov/sites/production/files/2017-12/documents/ffrro_factsheet_pbb_11-16-17_508.pdf [accessed 2020-01-23]
20. Croup: Symptoms and Causes. Mayo Clinic. 2019 Apr 11. URL: https://www.mayoclinic.org/diseases-conditions/croup/symptoms-causes/syc-20350348 [accessed 2020-01-23]
21. Biggest open problems in natural language processing. Sciforce. 2020 Feb 05. URL: https://medium.com/sciforce/biggest-open-problems-in-natural-language-processing-7eb101ccfc9 [accessed 2020-01-23]
22. Zhang XA, Yates A, Vasilevsky N, Gourdine JP, Callahan TJ, Carmody LC, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit Med 2019;2 [FREE Full text] [doi: 10.1038/s41746-019-0110-4] [Medline: 31119199]
23. Smirnov A, Levashova T. Knowledge fusion patterns: A survey. Information Fusion 2019 Dec;52:31-40. [doi: 10.1016/j.inffus.2018.11.007]
24. TranQL. URL: https://tranql.renci.org [accessed 2020-11-02]
25. NCATS-Tangerine / tranql. URL: https://github.com/NCATS-Tangerine/tranql [accessed 2020-11-02]
26. TranQL API. URL: https://tranql.renci.org/apidocs/ [accessed 2020-11-02]
27. What is Translator? Translator Program: Teams Green and Gamma. URL: https://researchsoftwareinstitute.github.io/data-translator/ [accessed 2020-11-02]

Abbreviations

2D: 2-dimensional
API: application programming interface
GPU: graphical processing unit
ICEES: Integrated Clinical and Environmental Exposures Service
IYD: iodotyrosine deiodinase
KG: knowledge graph
KGS: knowledge graph standard
PTH: parathyroid hormone
ROBOKOP: Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways
TranQL: Translator Query Language
UI: user interface

Edited by C Lovis; submitted 24.01.20; peer-reviewed by A Aminbeidokhti, J Yang, N Mohammad Gholi Mezerji; comments to author 06.05.20; revised version received 30.06.20; accepted 17.07.20; published 23.11.20.

Please cite as:
Cox S, Ahalt SC, Balhoff J, Bizon C, Fecho K, Kebede Y, Morton K, Tropsha A, Wang P, Xu H
Visualization Environment for Federated Knowledge Graphs: Development of an Interactive Biomedical Query Language and Web Application Interface
JMIR Med Inform 2020;8(11):e17964
URL: http://medinform.jmir.org/2020/11/e17964/
doi:10.2196/17964
PMID:33226347

Original Paper

Use of Social Media by Hospitals and Clinics in Japan: Descriptive Study

Yuya Sugawara[1,2,3], PhD; Masayasu Murakami[2], PhD; Hiroto Narimatsu[3,4], MD, PhD

[1]Institute for Promotion of Medical Science Research, Faculty of Medicine, Yamagata University, Yamagata, Japan

[2]Department of Health Policy Science, Graduate School of Medical Science, Yamagata University, Yamagata, Japan

[3]Cancer Prevention and Control Division, Kanagawa Cancer Center Research Institute, Yokohama, Japan

[4]Graduate School of Health Innovation, Kanagawa University of Human Services, Kawasaki, Japan

Corresponding Author:
Hiroto Narimatsu, MD, PhD
Cancer Prevention and Control Division
Kanagawa Cancer Center Research Institute
2-3-2 Nakao Asahi-ku
Yokohama, 241-8515
Japan
Phone: 81 45 520 2222
Email: hiroto-narimatsu@umin.org

Abstract

Background: The use of social media by hospitals has become widespread in the United States and Western European countries. However, in Japan, the extent to which hospitals and clinics use social media is unknown. Furthermore, recent revisions to the Medical Care Act may subject social media content to regulation.

Objective: The purpose of this study was to examine social media use in Japanese hospitals and clinics. We investigated the adoption of social media, analyzed social media content, and compared content with medical advertising guidelines.

Methods: We randomly sampled 300 hospitals and 300 clinics from a list of medical institutions that was compiled by the Ministry of Health, Labour and Welfare. We performed web and social media (Facebook and Twitter) searches using the hospital and clinic names to determine whether they had social media accounts. We collected Facebook posts and Twitter tweets and categorized them based on their content (eg, health promotion, participation in academic meetings and publications, public relations or news announcements, and recruitment). We compared the collected content with medical advertising guidelines.

Results: We found that 26.0% (78/300) of the hospitals and 7.7% (23/300) of the clinics used Facebook, Twitter, or both. Public relations or news announcements accounted for 53.99% (724/1341) of the Facebook posts by hospitals and 58.4% (122/209) of the Facebook posts by clinics. In hospitals, 16/1341 (1.19%) Facebook posts and 6/574 (1.0%) tweets and in clinics, 8/209 (3.8%) Facebook posts and 15/330 (4.5%) tweets could conflict medical advertising guidelines.

Conclusions: Fewer hospitals and clinics in Japan use social media as compared to other countries. Social media were mainly used for public relations. Some content disseminated by medical institutions could conflict with medical advertising guidelines. This study may serve as a reference for medical institutions to guide social media usage and may help improve medical website advertising in Japan.

(JMIR Med Inform 2020;8(11):e18666) doi:10.2196/18666

KEYWORDS

social media; internet; hospitals; health promotion; Japan

Introduction

More than 4.5 billion people use the internet, and the number of social media users worldwide has passed the 3.8 billion mark as of the start of 2020 [1]. Facebook and Twitter are popular social media tools. As of the second quarter of 2020, Facebook had over 2.7 billion monthly active users (MAUs) [2]. As of the first quarter of 2019, Twitter had an average of 330 million MAUs worldwide [3]. In Japan, in 2019, Facebook had 26 million MAUs, and Twitter had about 48 million MAUs [4].

A major benefit of social media for health communication is the accessibility and widening access of health information to

various population groups, regardless of age, education, race or ethnicity, and locality [5]. Thus, many health care organizations use social media. In the United States, 94.41% of hospitals have Facebook pages, and 50.82% have Twitter accounts [6]. In Western Europe, 67.0% of hospitals have Facebook pages, and 18.1% have Twitter accounts [7]. Thaker et al [8] reported that hospitals use social media to announce news and events and to promote themselves and health.

While many hospitals disseminate beneficial health information, there is concern that some hospital social media content may breach patient privacy [9]. Some hospitals may disseminate blatant advertising [10]. Some plastic surgeons emphasize immediate positive results, without discussing any potential complications or postoperative care requirements, and photoshopped before and after pictures are commonplace in some social media posts [10]. Japan's Medical Care Act was amended in 2017 and the new Medical Care Act has been enforced since 2018 [11,12]. With this revision, websites of medical institutions that were previously not considered advertisements are now considered so and are also now subject to regulation. To this end, administrative and criminal penalties have been introduced for violations. In addition, medical advertisement guidelines were also revised by the Ministry of Health, Labour and Welfare (MHLW) [13]. The contents of health care organizations' websites have been restricted because of these revisions. Although the guidelines do not mention social media, they may be identified as websites, and the contents of health care organizations' social media can thus be restricted. However, Japanese hospitals and clinics may disseminate health information that do not follow medical advertisement guidelines.

Presently, the extent to which Japanese health care organizations use social media is unknown, necessitating investigation. Accordingly, this study was designed to investigate the outline of social media use in Japanese hospitals and clinics through the following research questions:

1. How many social media accounts do Japanese medical institutions have?

2. What kind of information do Japanese medical institutions post on social media?

3. Does the information posted by the medical institutions conform to the medical advertising guidelines?

Methods

Study Population

We extracted study samples based on lists that were available. The list of insurance-covered medical institutions is maintained by the Regional Bureau of Health and Welfare, MHLW [14,15]. The list of clinics that performed treatment not covered by health insurance was published on Yahoo! Healthcare [16]. We extracted 8600 hospitals, 154,213 clinics, and 515 clinics that performed treatment not covered by health insurance, from the lists of medical institutions.

In Japan, the universal insurance system was established in 1961 [17]. This system allows anyone to visit medical institutions anytime and anywhere with no discrimination [18,19]. Therefore, we assumed that there was no difference in regional medical care provision and the use of social media. In this study, 300 samples were uniformly extracted from hospitals and clinics in Japan without considering regional bias. We assigned a random number to each hospital and clinic using an Excel (Microsoft) function. After assigning a random number, 300 samples were extracted in the descending order of random numbers. The size of the extracted sample was estimated based on the interval estimation of the population proportion. We performed a pilot study from February 23, 2018, to March 12, 2018, extracting 200 samples for trial. The results indicated that 26.5% (53/200) of hospitals used social media. We estimated that the sample size was 300 by using the statistical software EZR with the width of the 95% confidence interval as 0.1, so that the actual results fit within ±5% of the true value with the expected proportion being 0.265. Moreover, the 200 test samples were not included in the 300 samples used in the main study.

The date of designation as insurance-covered medical institutions, the name of the medical institutions, address, phone number, ID of medical institutions, and specialty are contained in the list of insurance-covered medical institutions. This list has been published on a website maintained the Regional Bureau Health and Welfare of each region [20-27]. Anyone can freely download the list as a PDF file (Adobe) or MS Excel file. We used the data of insurance-covered medical institutions as of October 1, 2017, in this study. We accessed Yahoo! Healthcare to collect the data not covered by health insurance clinics on November 5, 2017. However, Yahoo! Healthcare, which published information on health care and medical institutions on its website, was shut down on March 29, 2018 [16].

Social Media Accounts of Hospitals and Clinics

Facebook and Twitter, the major social media in Japan, were selected for analysis. For each sample of 300 hospitals and clinics, we performed Google searches using the names of the hospitals and clinics. We checked whether social media accounts of hospitals and clinics exist. Using the search engine on the official social media page as well as Google, the name of each medical institution was searched to check for the existence of a social media account. For medical institutions that have social media accounts, their websites were checked to see whether a social media policy has been formulated.

We surveyed the numbers of "likes" and "followers" from the medical institutions' Facebook and Twitter pages. The attributes of each hospital and clinic (clinical department, number of beds, types of beds, who established it) were drawn from the extracted hospitals and clinics' websites and the list of insurance-covered medical institutions.

The survey of the social media accounts of hospitals and clinics was conducted from April 7 to April 22, 2018. We accessed social media accounts of hospitals from April 7 to 15, 2018, and clinics from April 15 to 19, 2018. The data gathering of social media accounts was completed on April 22, 2018.

Data Collection

We collected content from Facebook and Twitter. For each hospital and clinic account, we collected 100 Facebook posts and 1000 tweets. Content data were collected using NodeXL

Excel Template 2014 (version 1.0.1.402; The Social Media Research Foundation), an MS Excel add-in [28]. After collecting Facebook posts and tweets, to investigate the difference in the number of comments for each season, the number of monthly comments in 2017 for each hospital and clinic was calculated. Then, Facebook post and Twitter tweet data were collected between August 4 and 5, 2018.

Classification of Contents

The latest 20 Facebook posts and the latest 100 tweets were manually categorized by content per medical institution. Contents were categorized manually and classified into 4 types: "Health promotion," "Participation in academic meetings, publications," "Public relations, news announcements," and "Recruitment." At first, it was divided into "Health promotion," "Public relations, news announcements," and "Recruitment" with reference to previous studies [5,8,29,30]. As we continued the classification, we found that there was a lot of content related to participation in academic meetings and publications. Therefore, a new item "Participation in academic meetings, publications" was added. We categorized social media contents as shown in Textbox 1.

If the social media content was updates on the medical institution's blog, we checked the links and categorized the comments. If more than 1 content is included, the main topic is judged from the context and the comments are categorized.

Three researchers (a medical informatics specialist [YS], a health policy specialist [MM], a medical doctor and public health specialist [HN]) categorized contents into 4 types. When a conflict occurred, it was resolved by discussions between the 3 researchers. Thus, all content was categorized upon agreement from the 3 researchers.

Textbox 1. Social media contents.

- Health Promotion: Dissemination of medical knowledge and health information. This includes easy-to-understand medical knowledge and health information for patients and the public, and professional information for professionals.
- Participation in Academic Meetings, Publications: Comments on academic activities such as information on holding academic meetings, participation in academic meetings, writing papers, and specialized books.
- Public Relations, News Announcements: Reports on in-hospital events for patients, notifications from hospitals, public relations, comments related to consultation (eg, hospitals are closed, change in consultation hours).
- Recruitment: Content related to human resources, such as personnel change reports and comments on recruitment.
- Others: Comments that do not apply to any of the above. For example, comments on activities that are not related to the actual work, such as welcome parties, social gatherings, and sports competitions.

Comparison With Guidelines

We compared the collected contents with the medical advertising guidelines and examined whether they complied with the guidelines or were appropriate as advertisements. In addition to the medical advertising guidelines, the "Doctor's Professional Ethics Guidelines," "The way medical facility websites should be – Guidelines for providing member medical facilities and medical information" (2008 March revised edition; both issued by the Japan Medical Association), and a previous study that compared medical advertising guidelines and the websites of medical institutions related to aesthetic medicine were used to create evaluation items and criteria (Multimedia Appendix 1) [31-33]. Referring to the criteria and the advertising example described in the medical advertising guidelines, 3 researchers (a medical informatics specialist [YS], a health policy specialist [MM], a medical doctor and public health specialist [HN]) compared contents and medical advertising guidelines. Based on the agreement of the 3 researchers, it was decided whether it was appropriate as a medical advertisement.

Text Mining

To complement manual content analysis, text mining was performed on Facebook posts and tweets of the hospitals and clinics, respectively. We calculated term frequency—which is the number of occurrences of each target word in an entire text—and created a co-occurrence network. We used KH Coder Version 3.Beta.01g for Windows for this task [34-36]. ChaSen, which was used for the morphological analysis, was included in KH Coder and used for word extraction. KH Coder uses the Jaccard coefficient to determine the degree of word-to-word co-occurrence and creates a network chart [37]. In this chart, words closely associated with each other are connected with lines [37]. KH Coder also displays networks that are more closely associated with each other as "subgraphs" through color coding [37]. In this context, co-occurrence means there is a close relationship between words [38].

Statistical Analysis

The percentage of the social media account holding ratio for each medical institution was calculated. We regarded a medical institution that has either or both a Facebook and Twitter account as "Having a social media account." We calculated the median and IQR for the numbers of beds, Facebook likes, and Twitter followers.

Fisher exact test and logistic regression analysis were performed on the attributes of medical institutions and whether medical institutions have social media accounts.

Hospital attributes were hospital size (small and medium hospitals with 20-199 beds, large hospitals with more than 200 beds), urban/rural, hospital classification (general hospital, internal medicine hospital, surgical hospital), who established it (individual/nonprofit medical corporations, national/public/social insurance-related organizations), Regional Bureau of Health and Welfare in each region, hospital functions (general hospitals, special functioning hospitals or regional

medical care support hospitals), and whether the hospital has a website.

Clinic attributes were whether the clinic has a bed, urban/rural, medical/dental classification, Regional Bureau of Health and Welfare in each region, who established it (individuals, nonprofit medical corporations, national/public), specialty (internal medicine departments, surgical departments, dentistry), and whether the clinic has a website.

In the logistic regression analysis, the presence or absence of social media accounts was analyzed as a dependent variable, and the attributes of medical institutions were analyzed as independent variables.

We compared the ratio of sample medical institutions by region with actual medical institutions. The goodness-of-fit test was performed by the chi-square test with reference to the reports released by the MHLW [39].

A *P* value <.05 was considered statistically significant. Statistical analyses were performed with EZR (version 1.37, Saitama Medical Center, Jichi Medical University), which is a graphical user interface for R (The R Foundation for Statistical Computing). More precisely, it is a modified version of R commander designed to add statistical functions frequently used in biostatistics [40].

This study was approved by the Institutional Review Board of Yamagata University, Faculty of Medicine.

Results

Sample Medical Institutions

We extracted 600 medical institutions (300 hospitals and 300 clinics). Of the 300 hospitals, 209 were small and medium hospitals, and 91 were large hospitals; 10 hospitals were special functioning hospitals or regional medical care support hospitals. Of the 300 clinics, 176 were medical clinics and 124 were dental clinics. For the ratio of number of sample medical institutions to the actual number of medical institutions by each Regional Bureau of Health and Welfare, a chi-square test revealed no significant difference in hospitals (P=.268, χ^2_7=8.791) or clinics (P=.958, χ^2_7=2.028). Multimedia Appendix 2 shows a table comparing the ratio of sample medical institutions to actual medical institutions.

Research Question 1

Hospital Accounts

Table 1 shows the number and ownership of social media accounts of medical institutions. Of the 300 hospitals and clinics, 78 (26.0%) and 23 (7.7%), respectively, have Facebook or Twitter accounts or both.

Tables 2 and 3 show the results of Fisher exact test and logistic regression analysis for the use of social media and the attributes of hospitals, respectively. The Fisher exact test showed a significant difference in the presence or absence of social media and hospital size (P<.001), hospital classification (P=.018), hospital function (P=.004), and website presence (P=.025). Logistic regression analysis showed a significant difference in hospital size (P<.001). The odds ratio was 3.25 with a 95% confidence interval ranging from 1.75 to 6.04. No significant difference was found except for hospital size. The ranges of all generalized variance inflation factor in the logistic regression analysis ranged from 1.00 to 1.39.

Table 1. Numbers and percentages of social media accounts and websites that medical institutions had (N=300).

Institutions	Social media	Facebook	Twitter	Website
Hospitals, n (%)	78 (26.0)	73 (24.3)	13 (4.3)	286 (95.3)
Clinics, n (%)	23 (7.7)	19 (6.3)	11 (3.7)	129 (43.0)

Table 2. Fisher exact test regarding the use of social media and the attributes of medical institutions (hospitals).

Item and Classification	Not using social media (N=222)	Using social media (N=78)	*P* value
Hospital size			**<.001**
Small and medium hospitals with 20-199 beds, n (%)	171 (77.0)	38 (48.7)	
Large hospitals with more than 200 beds, n (%)	51 (23.0)	40 (51.3)	
Urban, rural			**.364**
Rural, n (%)	18 (8.1)	9 (11.5)	
Urban, n (%)	204 (91.9)	69 (88.5)	
Hospital classification			**.018**
General hospital, n (%)	108 (48.6)	51 (65.4)	
Internal medicine hospital, n (%)	99 (44.6)	21 (26.9)	
Surgical hospital, n (%)	15 (6.8)	6 (7.7)	
Established by			**.073**
Individual/nonprofit medical corporations, n (%)	182 (82.0)	56 (71.8)	
National/public/social insurance-related organizations, n (%)	40 (18.0)	22 (28.2)	
Regional Bureau of Health and Welfare			**.233**
Hokkaido, n (%)	22 (9.9)	4 (5.1)	
Tohoku, n (%)	15 (6.8)	6 (7.7)	
Kanto-Shinetsu, n (%)	48 (21.6)	28 (35.9)	
Tokai-Hokuriku, n (%)	26 (11.7)	8 (10.3)	
Kinki, n (%)	43 (19.4)	10 (12.8)	
Chugoku-Shikoku, n (%)	21 (9.5)	10 (12.8)	
Shikoku, n (%)	9 (4.1)	3 (3.8)	
Kyushu, n (%)	38 (17.1)	9 (11.5)	
Hospital function			**.004**
General hospital, n (%)	219 (98.6)	71 (91.0)	
Special functioning hospitals or regional medical care support hospitals, n (%)	3 (1.4)	7 (9.0)	
Website			**.025**
Absent, n (%)	14 (6.3)	0 (0.0)	
Present, n (%)	208 (93.7)	78 (100.0)	
Beds, median (IQR)	120.00 (69.25-198.75)	220.00 (100.50-370.00)	<.001
Facebook likes, median (IQR)	N/A[a]	66.00 (19.00-207.00)	
Twitter followers, median (IQR)	N/A	7.00 (3.00-84.00)	

[a]NA: not applicable.

Table 3. Logistic regression analysis of hospital attributes and social media usage (hospitals).

Variable	Odds ratio (95% confidence interval)	*P* value
Hospital size		
Small and medium hospitals with 20 to 199 beds	Reference	<.001
Large hospitals with more than 200 beds	3.25 (1.75-6.04)	
Hospital classification		
General hospital	Reference	
Internal medicine hospital	0.57 (0.30-1.09)	.088
Surgical hospital	1.49 (0.51-4.32)	.46
Established by		
Individual/nonprofit medical corporations	Reference	
National/public/social insurance-related organizations	0.73 (0.34-1.56)	.41
Hospital function		
General hospital	Reference	
Special functioning hospitals or regional medical care support hospitals	3.27 (0.75-14.40)	.12
Website		
Absent	Reference	
Present	9200000.00 (0.00-infinity)	.99

Clinic Accounts

The number of Facebook and Twitter accounts of clinics was 19/300 (6.3%) and 11/300 (3.7%), respectively (Table 1). Tables 4 and 5 show the results of Fisher exact test and logistic regression analysis. The Fisher test showed a significant difference in website (P<.001). Logistic regression analysis showed a significant difference in website (P<.001) and specialty (dentistry, P=.037). Generalized variance inflation factor in logistic regression analysis was 1.01.

Table 4. Fisher exact test regarding the use of social media and the attributes of medical institutions (clinics).

Item and classification	Not using social media (N=277)	Using social media (N=23)	*P* value
Beds			**.637**
Absent, n (%)	261 (94.2)	21 (91.3)	
Present, n (%)	16 (5.8)	2 (8.7)	
Urban/Rural			**.615**
Rural, n (%)	15 (5.4)	0 (0.0)	
Urban, n (%)	262 (94.6)	23 (100.0)	
Medical/Dental classification			**.13**
Dental clinics, n (%)	111 (40.1)	13 (56.5)	
Medical clinics, n (%)	166 (59.9)	10 (43.5)	
Regional Bureau of Health and Welfare			**.408**
Hokkaido, n (%)	7 (2.5)	2 (8.7)	
Tohoku, n (%)	18 (6.5)	0 (0.0)	
Kanto-Shinetsu, n (%)	110 (39.7)	9 (39.1)	
Tokai-Hokuriku, n (%)	36 (13.0)	1 (4.3)	
Kinki, n (%)	46 (16.6)	4 (17.4)	
Chugoku-Shikoku, n (%)	15 (5.4)	2 (8.7)	
Shikoku, n (%)	10 (3.6)	1 (4.3)	
Kyushu, n (%)	35 (12.6)	4 (17.4)	
Established by			**.736**
Individual, n (%)	167 (60.3)	13 (56.5)	
Nonprofit medical corporations, n (%)	107 (38.6)	10 (43.5)	
National/public/social insurance related organizations, n (%)	3 (1.1)	0 (0.0)	
Specialty			**.185**
Internal medicine departments, n (%)	95 (34.3)	4 (17.4)	
Surgical departments, n (%)	71 (25.6)	6 (26.1)	
Dentistry, n (%)	111 (40.1)	13 (56.5)	
Website			**<.001**
Absent, n (%)	169 (61.0)	2 (8.7)	
Present, n (%)	108 (39.0)	21 (91.3)	
Beds, median (IQR)	0.00 (0.00-0.00)	0.00 (0.00-0.00)	.57
Facebook likes, median (IQR)	N/A[a]	69.00 (24.50-95.50)	
Twitter follower, median (IQR)	N/A	11.00 (3.00-23.50)	

[a]NA: not applicable.

Table 5. Logistic regression analysis of hospital attributes and social media usage (clinics).

Variable	Odds ratio (95% confidence interval)	*P* value
Website		
Absent	Reference	
Present	17.80 (4.07-78.20)	<.001
Specialty		
Internal medicine departments	Reference	
Surgical departments	2.20 (0.58-8.40)	.25
Dentistry	3.55 (1.08-11.70)	.037

Social Media Policy

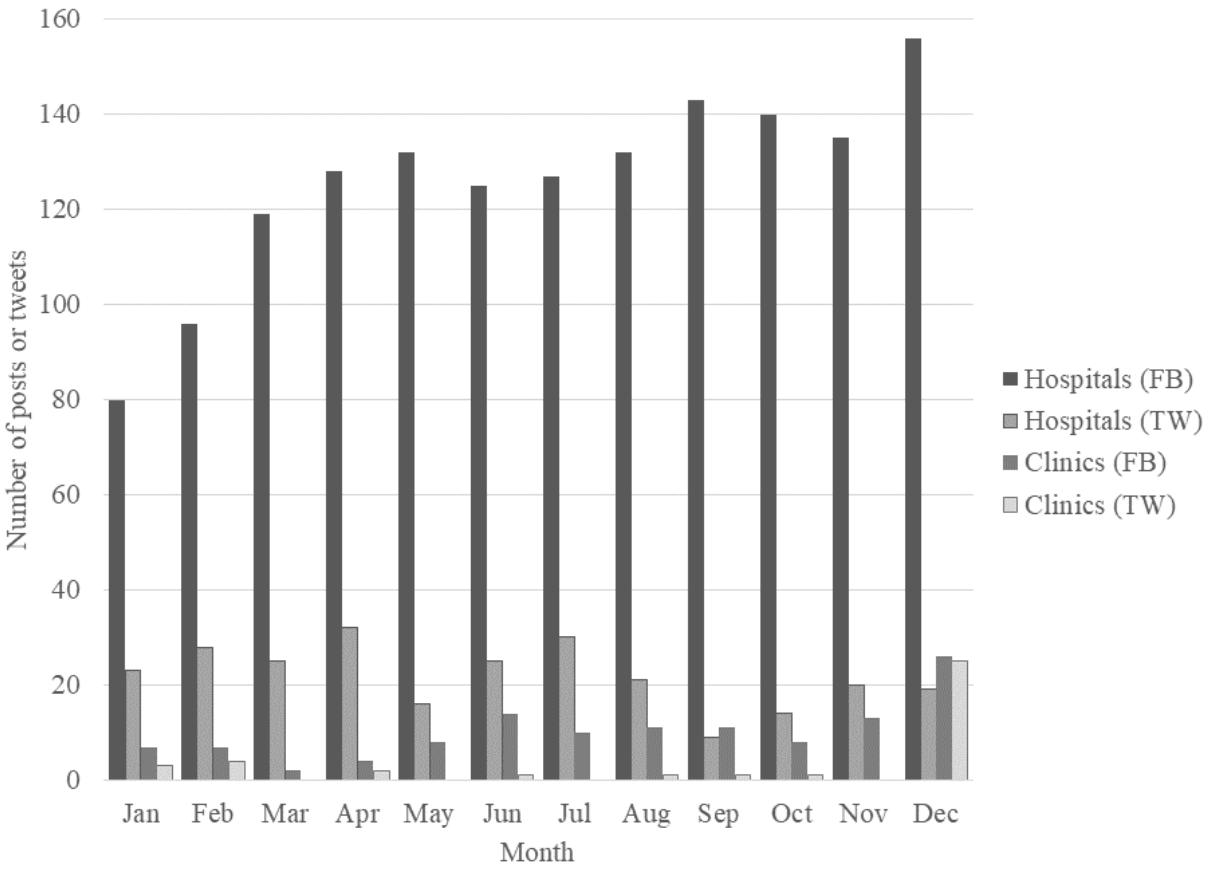
Three hospitals and no clinics disclosed social media usage policies on their website.

Research Question 2

Number of Comments

The total number of social media messages disseminated by medical institutions was 8026 from September 16, 2010, to August 4, 2018. Hospitals published 4514 Facebook posts and 2679 tweets, whereas clinics had 503 Facebook posts and 330 tweets. The number of comments we used for content analysis was 1341 hospital Facebook comments and 574 Twitter comments; for clinics, 209 Facebook comments, and 330 Twitter comments. Figure 1 shows the number of monthly comments for the year 2017. For both hospitals and clinics, Facebook posts and tweets all increased in December. In 2017, the annual number of comments for hospitals was 1513 Facebook comments and 262 Twitter comments; for clinics, 121 Facebook comments and 38 Twitter comments. Multimedia Appendix 3 shows examples of contents of hospitals and clinics.

Figure 1. The number of comments for hospitals and clinics in 2017. The number of comments from the clinic was small. The number of comments increased in December at both hospitals and clinics. FB: Facebook; TW: Twitter.



Classification of Contents

Figures 2 and 3 show the classification of Facebook and Twitter content of hospitals and clinics, respectively. For hospitals and clinics, "Public relations, news announcement" was the highest, accounting for more than 50% of the content (hospital Facebook content: 53.99% [724/1341]; hospital Twitter content: 66.6% [382/574]; clinic Facebook content: 58.4% [122/209]; clinic Twitter content: 56.4% [186/330]). Compared to hospitals, clinics had posted more "Health promotion" tweets on Twitter. For hospitals using Facebook, "Participation in academic meetings, publications" accounted for 24.09% (323/1341) of the posts, but few in hospitals using Twitter and clinics. Hospitals and clinics disseminated little content related to "Recruitment" on Facebook and Twitter.

Figure 2. Classification and percentage of social media messages (Hospitals). The latest 20 Facebook posts and the latest 100 tweets were manually categorized by content per medical institution. "Participation in academic meetings, publications" accounted for 24.1% of the Facebook posts.

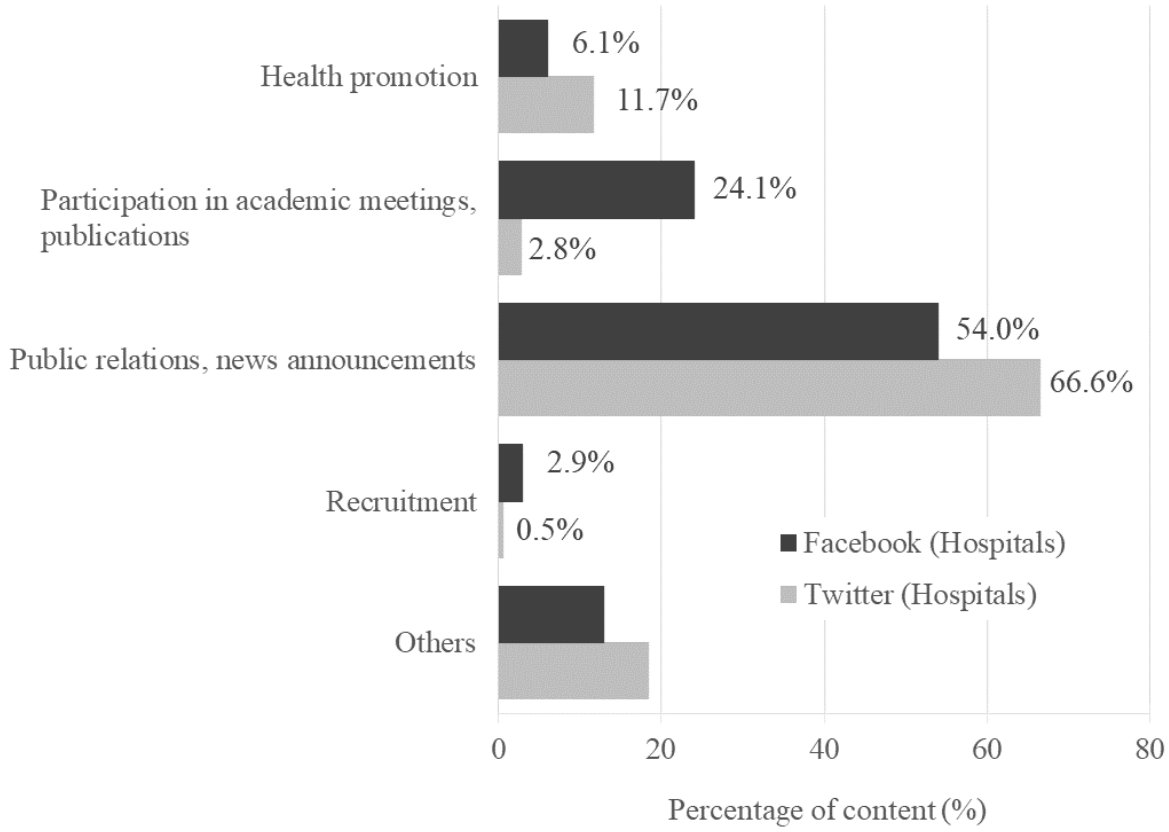
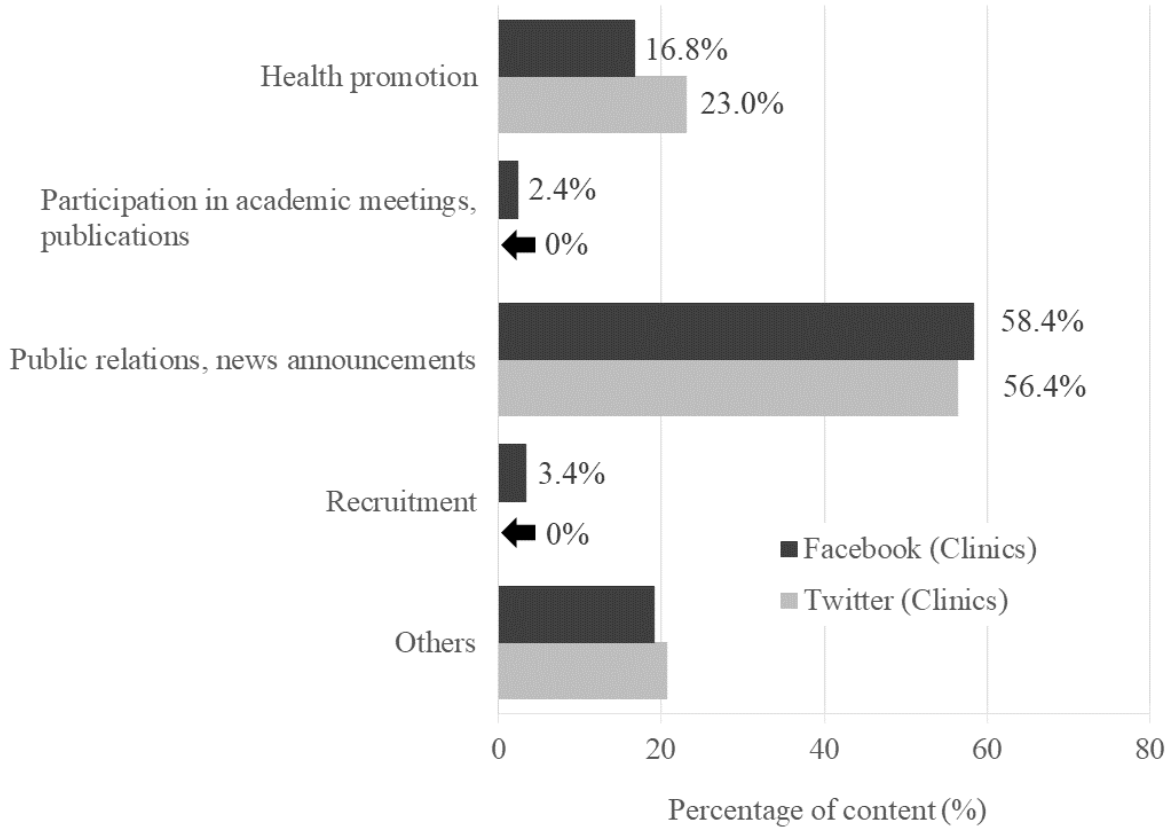


Figure 3. Classification and percentage of social media messages (clinics). The latest 20 Facebook posts and the latest 100 tweets were manually categorized by content per medical institution. Higher percentage of "Health promotion" compared to hospitals.



Term Frequency and Co-occurrence Network

The results of text mining are shown in Figures 4 and 5, and Multimedia Appendix 4. On the Facebook accounts of hospitals, more words related to conference presentations appeared than others. The frequency was 815 times for "academic meeting," 746 times for "presentation," and 635 times for "research," thus, forming a co-occurrence network. On hospital Twitter accounts, "influenza" formed a co-occurrence network. At clinics, there were many announcements about leave of absence on both Facebook and Twitter. The number of occurrences of "closed" was 158 and 73, respectively, on Facebook and Twitter.

Figure 4. A co-occurrence network for the hospitals in this study. Words in the same subgraph are connected by a solid line. When co-occurring with words in other subgraphs, they are connected by a broken line. Information related to nursing care, community-based health care, academic meeting, and lectures was posted on Facebook. On Twitter, there were tweets about a fun party at a hospital and tweets about updating the blog of hospital B.

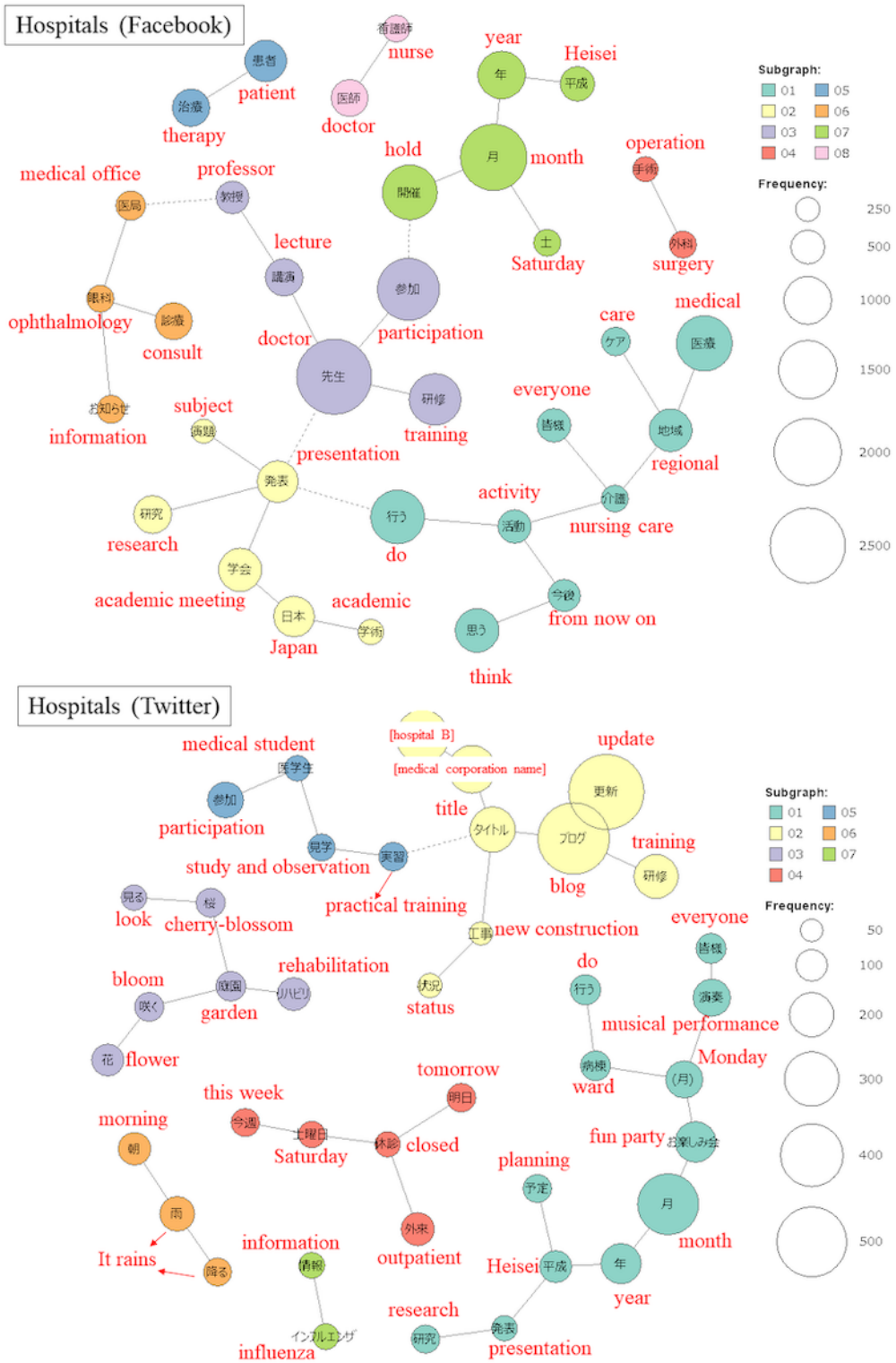
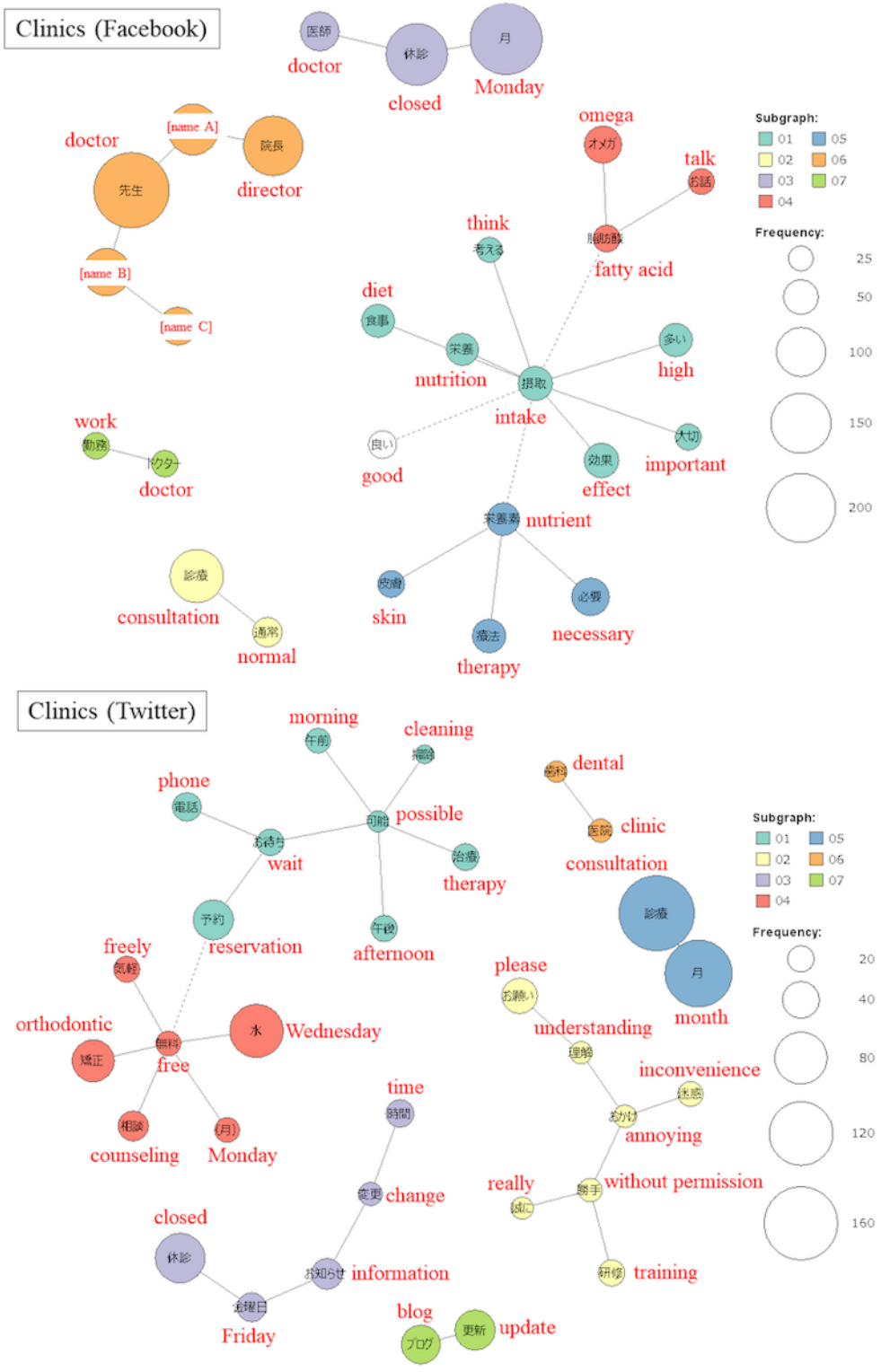


Figure 5. A co-occurrence network for the clinics in this study. Words in the same subgraph are connected by a solid line. When co-occurring with words in other subgraphs, they are connected by a broken line. Single words that do not belong to any subgraph are shown in white. On Facebook, a clinic was raising awareness about nutrition and omega fatty acids. On Twitter, there were tweets about free counseling on orthodontics.



Research Question 3

Table 6 shows the comparison between social media contents and guidelines. Content that could conflict with the guidelines and the percentage of total content by hospitals using Facebook and Twitter were 16 (1.19%, 16/1341) and 6 (1.0%, 6/574), respectively. In clinics, 8 Facebook posts (3.8%, 8/209) and 15 tweets (4.5%, 15/330) could conflict medical advertising guidelines. Multimedia Appendix 5 shows examples of this content.

Table 6. Number of messages that may violate medical advertising guidelines and professional ethics.

Evaluation items[a]	Applicable part of the guidelines[b]	Hospitals		Clinics	
		Facebook	Twitter	Facebook	Twitter
Introduction in media	By quoting or publishing articles in newspapers and magazines, discourses, theories, and experiences of doctors and scholars	12	4	4	0
Messages on safety	Misleading advertising	0	0	1	0
Invitation by matters not related to providing medical care	Advertising that impairs dignity	0	0	0	5
Emphasis on cost	Advertising that impairs dignity	0	2	1	10
Medical department name	Not included in advertisable items	0	0	0	0
Professional qualification	Not included in advertisable items	1	0	0	0
Regulations by other laws and regulations	Advertising prohibited by other laws or other advertising guidelines	1	0	0	0
Messages suggesting the superiority of the medical institution by comparison, exaggerated expressions of facility size, staffing, or medical provision	Advertising that implies their superiority by comparison/Misleading advertising	0	0	2	0
Ethical issues	Refer to the "Doctors' Professional Ethics Guidelines" issued by the Japan Medical Association	2	0	0	0

[a,b]Refer to Multimedia Appendix 1.

Discussion

Preliminary Findings

In this study, 300 hospitals and clinics, respectively, were sampled and classified according to social media accounts and their contents. In Japan, fewer medical institutions use social media than those in the United States and Western Europe. In addition, medical institutions using social media frequently used them as part of public relations activities. Some included messages that may violate medical advertising guidelines. To protect the reputation of medical institutions, it is considered necessary to formulate social media policies.

Social Media Accounts

In Japan, social media were rarely used by medical institutions, and it was considered that websites were mainly used for the dissemination of health information by medical institutions (Table 1). An online survey on health awareness among 3000 people showed that less than 5% used social networking sites as health information sources [41]. For this reason, even if a medical institution creates a social media account, only few users possibly refer to social media information from medical institutions. Because of a limited number of users, the number of "likes" and "followers" would not increase, and it would be difficult for medical institutions to ascertain the influence of using social media. As a result, medical institutions will interrupt the use of social media. In the United States, social media are an important source of information for using health information on the internet. According to a survey conducted in the United States in 2011, about one-fifth of approximately 23,000 respondents said that social media were the source of health information. In addition, one-third of respondents reported that social media are a reliable information source [42]. This viewpoint difference about social media between Japan and the United States may be reflected in the differences in social media utilization rates by medical institutions.

Social Media Utilization and Benefits

Social media have been used to maintain or improve peer-to-peer and clinician-to-patient communication, promote institutional branding, and improve the speed of interaction between and across different health care stakeholders in the health care field [43]. Patients may perceive that hospitals with social media activity are likely to offer advanced technologies and cutting-edge therapies [6].

In Japan, more than 50% of the social media comments sent by medical institutions were related to public relations activities. About a quarter of Facebook posts by hospitals were related to participation in academic conferences and the publication of academic papers (Figure 2). In text mining the Facebook accounts of hospitals, the frequency of "academic meeting" and "presentation" was high (Multimedia Appendix 4). Subgraphs related to conference presentations also appeared in the co-occurrence network (Figure 4). In particular, hospitals may have used social media to disseminate academic information. Additionally, an apology posted on Facebook by a hospital regarding the emergency discharge (quenching) of helium gas from a magnetic resonance imaging system was found (Multimedia Appendix 3). In this context, several reports have claimed that social media are a useful communication tool in emergency situations, such as disasters and accidents [44-51]. Further, social media may be useful when we want to share information urgently, because they have the advantage of immediacy compared to conventional media.

Social media have often been used for the purpose of health promotion and health education [5], and such health information may be used to improve public health as well. However, only a few medical knowledge and health information messages are disseminated by medical institutions in Japan. It may be even better to consider the season when health information is disseminated, as the number of monthly comments increased in December (Figure 1), a possible reason being the increased number of comments about Christmas as well as the year-end and New Year holidays. However, when medical institutions disseminate information on social media, it may be good to raise awareness not only about annual events but also about seasonal diseases. In fact, the spread of awareness on influenza vaccination using social media is common [52]. In addition, information on pollen allergy is provided using a mobile app [53].

Social media use by medical institutions involves mostly one-way communication, and few medical institutions respond to inquiries from the general public or patients via social media [8]. However, two-way communication with the general public and patients may meet patient needs that cannot be met through daily medical care and may thus help improve the provision of care [54,55].

Risks and Problems in Using Social Media

There are some problems with medical institutions using social media. These include concerns about patient privacy breaches, issues with the reliability and poor quality of information, and the obscuring of boundaries between health care professionals and patients [5,56].

When medical institutions disseminate information on social media, great care should be taken not to breach patient privacy as seemingly innocent comments can do so [57]. Even if the post does not contain a specific name, it may be possible to identify the patient by indirect information such as the name of the town where the patient lives, gender, or disease name [57]. Thus, medical institutions should be cautious when posting on social media, as these privacy breaches may occur unintentionally.

It is often difficult to tell who wrote health information on social media, which raises concerns regarding its accuracy and reliability [56]. Additionally, if medical institutions use social media, it will be necessary to clarify the boundaries between health care professionals and patients. Few doctors and medical institutions respond to “friend” requests from patients [56], but it is better to prescribe what to do when receiving “friend requests” in the social media policy in advance.

Moreover, when a medical institution uses social media, it may be necessary to create a social media policy not only to clarify the purpose of social media use but also to protect its reputation [56,58-60]. In this study, only 3 medical institutions disclosed their social media policies on their website. Thus, many medical institutions might not develop social media policies. In this context, damage to reputation and breach of patient privacy are matters of concern when medical institutions use social media [58]. Consequently, medical institutions should have clear objectives [59] and develop social media policies to avoid these risks.

Comparison With the Guidelines

In this survey, no content that violated patient privacy was extracted. However, some contents that could violate the guidelines were extracted. Of the hospital’s Facebook posts, 0.89% (12/1341) commented on being featured in the media. According to medical advertising guidelines, coverage announcements are also considered as advertising, and they are essentially restricted. Therefore, when sending information through social media, it would be necessary to refrain from commenting on whether their facility and staff are featured in newspapers, magazines, and other media. There was also a hospital Facebook account that sent company advertisements directly without disclosing conflicts of interest. This is considered ethically problematic. The Doctor’s Professional Ethics Guidelines stipulate that the relationship with medical providers should be appropriate [32]. In the website guidelines by the Japan Medical Association, “advertising by external sponsors” is listed as ineligible content [31]. Similarly, there are provisions regarding conflicts of interest in overseas guidelines; the British Medical Association social media usage guidelines require disclosure of conflicts of interest when doctors and medical students post information online [61].

Some clinics posted tweets emphasizing costs and matters not related to medical provision. An example is the toothbrush gift campaign when visiting the dental clinic, as well as discount campaigns such as medical checkups and whitening. In general, when a company uses a social medium for promotional purposes, coupons are often issued and discounts are announced on the social medium [62,63]. Therefore, if a medical institution uses social media like a company, it may be easy to disseminate messages on examinations and treatment fees and discounts. However, according to medical advertising guidelines, advertising that emphasizes costs is considered “Advertising that impairs dignity,” and such messages should not be disseminated. By disseminating such inappropriate messages, medical institutions not only receive a reprimand from health authorities but may also lose their good reputation.

The government should probably respond to messages on social media. In this study, referring to medical advertisement guidelines and the literature, we determined whether social media contents disseminated by medical institutions violated the guidelines. For some contents, it was difficult to determine whether they meet the guidelines. Governments might need to articulate the criteria for determining whether their contents are appropriate or inappropriate. The MHLW’s internet patrol and public notification regarding medical institutions’ websites are currently in execution [64,65]. In addition, it may be necessary to strengthen checks on inappropriate social media cases.

Limitations

Sampling Methods

In this study, we randomly assigned a number to the list of medical institutions in Japan and extracted 300 small samples for each hospital and clinic. Compared to the actual number of medical institutions, these samples showed no statistically

significant difference in the number of medical institutions by region, as presented in this study. In this study, regional bias may be possible, but it may be limited. However, the samples may not be representative of all Japanese medical institutions. These samples may be biased when examined in detail with prefectures and cities. Additionally, the characteristics and attributes of medical institutions may be biased. For a detailed study of social media usage in Japanese medical institutions in the future, it may be necessary to increase the sample size and reduce the confidence interval width. In addition, sampling methods such as stratified random sampling, cluster sampling, and multistage sampling should be used to obtain more representative samples [66].

Content Analysis

In this study, the classification of contents and the comparison with the medical advertising guidelines were made based on the consensus of 3 researchers. However, it did not preclude personal subjectivity; classification and comparison may be inconsistent, and objective evaluation will be necessary in future research. Further, measurement of intercoder reliability, which is fundamental and important in content analysis [67], is required for an objective evaluation. In addition, a thematic analysis approach such as topic modeling is required for objective categorization [68,69].

Factors Affecting Social Media Use in Medical Institutions

Regarding the use of social media by medical institutions, this study does not clarify the factors that led to the use of social media or the reasons why they were not used. Thus, the application of the unified theory of acceptance and use of technology and technology acceptance model may be necessary to examine the factors behind the use of social media in medical institutions [70].

Other Social Media

In this study, the target social media were limited to Facebook and Twitter. In future studies, it will be necessary to investigate the use of other platforms, such as blogs, wikis, LINE, and Instagram accounts of medical institutions. Blogs have been used since as early as 2004, and Wikipedia is often used in the medical community [43]. However, there are no reports of their usage at medical institutions in Japan, and the details remain unknown. LINE was developed in Japan [71], and its usage rate in Japan is high. According to a Ministry of Internal Affairs and Communications survey of 1500 people in 2016, Facebook usage was 32.3% and Twitter usage was 27.5%, whereas LINE usage was 67.0%, the highest [72]. In fact, it has been reported in a newspaper that a medical institution already uses LINE [73]. If medical institutions use LINE, messages pertaining to public relations and awareness activities may be more effectively distributed than via Facebook and Twitter. Instagram is a photo-sharing site that has been rapidly growing by the increasing number of users in recent years [74]. Medical institutions may be able to promote public relations activities by posting visually appealing images of them on Instagram. However, images that may violate medical advertising guidelines may be posted.

Lack of Cosmetic Surgery Clinics in the Sample

In this study, we investigated the actual use of social media by medical institutions throughout Japan but did not include cosmetic surgery clinics in the sample. Cosmetic surgery clinics might disseminate more advertisements than other specialties because many cosmetic surgeries are performed as part of free medical care, and the ratio of content may differ from this survey.

Necessity of a Longitudinal Study

The data presented in this study are cross-sectional at the time of the survey. Previous studies have shown that the use of social media by medical institutions has changed over time [7,75]. Therefore, in Japan, it will be necessary to observe social media usage by medical institutions over time.

Conclusions

Social media usage by Japanese medical institutions is lower than that in the United States and Western European countries, and these media are mainly used for messages related to public relations. Some social media contents posted by medical institutions could conflict with medical advertising guidelines. In addition, few medical institutions have established social media policies. Due to deviations in usage rates from overseas and the characteristics of social media, it is necessary to consider social media other than Facebook and Twitter. This study may serve as a reference for medical institutions to guide social media usage and help improve medical website advertising in Japan.

Acknowledgments

This work was supported by JSPS KAKENHI (Grant Number JP18H00507). We thank Editage for English language editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Criteria for comparing social media content and guidelines, professional ethics.
[DOCX File , 18 KB - medinform_v8i11e18666_app1.docx]

Multimedia Appendix 2
Percentage of sample medical institutions by each region and percentage of actual medical institutions.
[DOCX File , 21 KB - medinform_v8i11e18666_app2.docx]

Multimedia Appendix 3
Examples of social media contents.
[DOCX File , 18 KB - medinform_v8i11e18666_app3.docx]

Multimedia Appendix 4
Number of frequencies of words in Facebook posts and tweets by hospitals and clinics top 50.
[DOCX File , 33 KB - medinform_v8i11e18666_app4.docx]

Multimedia Appendix 5
Examples of messages that may violate medical advertising guidelines and professional ethics.
[DOCX File , 17 KB - medinform_v8i11e18666_app5.docx]

References

1. Kemp S. Digital 2020: Global Digital Overview. 2020. URL: https://datareportal.com/reports/digital-2020-global-digital-overview [accessed 2020-10-20]
2. Statistita. Number of monthly active Facebook users worldwide as of 2nd quarter 2020 (in millions). URL: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ [accessed 2020-10-19]
3. Statistita. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions). URL: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ [accessed 2020-10-19]
4. Bugajski M. Japan's Top Social Media Networks for 2020. URL: https://www.humblebunny.com/japans-top-social-media-networks/ [accessed 2020-10-19]
5. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
6. Griffis HM, Kilaru AS, Werner RM, Asch DA, Hershey JC, Hill S, et al. Use of social media across US hospitals: descriptive analysis of adoption and utilization. J Med Internet Res 2014;16(11):e264 [FREE Full text] [doi: 10.2196/jmir.3758] [Medline: 25431831]
7. Van de Belt TH, Berben SAA, Samsom M, Engelen LJLPG, Schoonhoven L. Use of social media by Western European hospitals: longitudinal study. J Med Internet Res 2012;14(3):e61 [FREE Full text] [doi: 10.2196/jmir.1992] [Medline: 22549016]
8. Thaker SI, Nowacki AS, Mehta NB, Edwards AR. How U.S. hospitals use social media. Ann Intern Med 2011 May 17;154(10):707-708. [doi: 10.7326/0003-4819-154-10-201105170-00021] [Medline: 21576547]
9. Suby C. Social Media in Health Care: Benefits, Concerns, and Guidelines for Use. Creat Nurs 2013 Jan 01;19(3):140-147. [doi: 10.1891/1078-4535.19.3.140] [Medline: 24400468]
10. Reissis D, Shiatis A, Nikkhah D. Advertising on Social Media: The Plastic Surgeon's Prerogative. Aesthet Surg J 2017 Jan;37(1):NP1-NP2. [doi: 10.1093/asj/sjw174] [Medline: 27771608]
11. Health and Global Policy Institute. Overview of Major Legislation (The eighth revision to the 1948 Medical Care Act). URL: http://japanhpn.org/en/section-1-3/ [accessed 2020-10-19]
12. Sei S. Notes on websites due to Medical Care Act revisions (viewpoints for responding to Medical Care Act revisions and creating homepages) [In Japanese]. Byoin-Rashinban 2017/10/15 2017;8(112):18-22.
13. Ministry of Health, Labour and Welfare. Guidelines (Medical Advertising Guidelines) for Advertising for Medical and Dental Services , Hospitals and Clinics [In Japanese]. 2018. URL: https://www.mhlw.go.jp/file/06-Seisakujouhou-10800000-Iseikyoku/0000206548.pdf [accessed 2020-10-19]
14. Okamoto E. Current status of receipt online and outlook for national database [In Japanese]. 2009 Mar 3. URL: http://www.resept.com/yakujiexpert.pdf [accessed 2020-10-19]
15. Ehara A. The shortest distance between the center of population of each municipality nationwide and the core pediatrics and regional pediatric centers [In Japanese]. The Journal of the Japan Pediatric Society 2016;120(10):1508-1513.
16. Yahoo! Japan. News of yahoo! Health care End [In Japanese]. URL: https://medical.yahoo.co.jp/ [accessed 2018-04-01] [WebCite Cache ID 6yLmbU0PN]
17. Kondo A, Shigeoka H. Effects of universal health insurance on health care utilization, and supply-side responses: Evidence from Japan. Journal of Public Economics 2013 Mar;99:1-23 [FREE Full text] [doi: 10.1016/j.jpubeco.2012.12.004]
18. Ikegami N, Campbell JC. Health care reform in Japan: the virtues of muddling through. Health Aff (Millwood) 1999;18(3):56-75. [doi: 10.1377/hlthaff.18.3.56] [Medline: 10388203]

19. Ito A. Logic, ethical problems, and status of consensus concerning free access to health care insurance system [In Japanese]. Policy and Practice Studies 2018;4(1):125-137.
20. Chugoku-Shikoku Regional Bureau of Health and Welfare. List of insurance medical institutions within the jurisdiction of Chugoku-Shikoku Regional Bureau of Health and Welfare [In Japanese]. URL: https://kouseikyoku.mhlw.go.jp/chugokushikoku/chousaka/iryoukikanshitei.html [accessed 2020-10-19]
21. Hokkaido Regional Bureau of Health and Welfare. List of insurance medical institutions within the jurisdiction of Hokkaido Regional Bureau of Health and Welfare [In Japanese]. URL: https://kouseikyoku.mhlw.go.jp/hokkaido/gyomu/gyomu/hoken_kikan/code_ichiran.html [accessed 2020-10-19]
22. Kanto-Shinetsu Regional Bureau of Health and Welfare. List of insurance medical institutions within the jurisdiction of Kanto-Shinetsu Regional Bureau of Health and Welfare [In Japanese]. URL: https://kouseikyoku.mhlw.go.jp/kantoshinetsu/chousa/shitei.html [accessed 2020-10-19]
23. Kinki Regional Bureau of Health and Welfare. List of insurance medical institutions within the jurisdiction of Kinki Regional Bureau of Health and Welfare [In Japanese]. URL: https://kouseikyoku.mhlw.go.jp/kinki/tyousa/shinkishitei.html [accessed 2020-10-19]
24. Kyushu Regional Bureau of Health and Welfare. List of insurance medical institutions within the jurisdiction of Kyushu Regional Bureau of Health and Welfare [In Japanese]. URL: https://kouseikyoku.mhlw.go.jp/kyushu/gyomu/gyomu/hoken_kikan/index.html [accessed 2020-10-19]
25. Shikoku Regional Bureau of Health and Welfare. List of insurance medical institutions within the jurisdiction of Shikoku Regional Bureau of Health and Welfare [In Japanese]. URL: https://kouseikyoku.mhlw.go.jp/shikoku/gyomu/gyomu/hoken_kikan/shitei/index.html [accessed 2020-10-19]
26. Tohoku Regional Bureau of Health and Welfare. List of insurance medical institutions within the jurisdiction of Tohoku Regional Bureau of Health and Welfare [In Japanese]. URL: https://kouseikyoku.mhlw.go.jp/tohoku/gyomu/gyomu/hoken_kikan/itiran.html [accessed 2020-10-19]
27. Tokai-Hokuriku Regional Bureau of Health and Welfare. List of insurance medical institutions within the jurisdiction of Tokai-Hokuriku Regional Bureau of Health and Welfare [In Japanese]. URL: https://kouseikyoku.mhlw.go.jp/tokaihokuriku/gyomu/gyomu/hoken_kikan/shitei.html [accessed 2020-10-19]
28. The Social Media Research Foundation. NodeXL Graph Gallery: About NodeXL. URL: http://nodexlgraphgallery.org/Pages/AboutNodeXL.aspx [accessed 2020-10-19]
29. Richter JP, Muhlestein DB, Wilks CEA. Social media: how hospitals use it, and opportunities for future use. J Healthc Manag 2014;59(6):447-460. [Medline: 25647968]
30. Gomes C, Coustasse A. Tweeting and Treating: How Hospitals Use Twitter to Improve Care. Health Care Manag (Frederick) 2015;34(3):203-214. [doi: 10.1097/HCM.0000000000000063] [Medline: 26217995]
31. Japan Medical Association. The way medical facility websites should be -Guidelines for providing member medical facilities and medical information- (2008 March revised edition) [In Japanese]. 2008 Mar. URL: http://dl.med.or.jp/dl-med/nichikara/hp_guide.pdf [accessed 2020-10-19]
32. Japan Medical Association. Doctor's Professional Ethics Guidelines Third edition October 2016 [In Japanese]. 2016. URL: http://dl.med.or.jp/dl-med/teireikaiken/20161012_2.pdf [accessed 2020-10-19]
33. Ohba H. Research on the Information Provision of the Aesthetic Medical Service in the Aesthetic Medical Institutions' Websites [In Japanese]. Journal of the Japan Association for Medical Informatics 2016;36(2):79-84.
34. Higuchi K. A Two-Step Approach to Quantitative Content Analysis: KH Coder Tutorial Using Anne of Green Gables (Part I). Ritsumeikan Social Science Review 2016;52(3):77-91.
35. Higuchi K. A Two-Step Approach to Quantitative Content Analysis: KH Coder Tutorial Using Anne of Green Gables (Part II). Ritsumeikan Social Science Review 2017;53(1):137-147.
36. Higuchi K. KH Coder Index Page. URL: https://khcoder.net/en/ [accessed 2020-10-19]
37. Higuchi K. KH Coder 3 Reference Manual. URL: https://khcoder.net/en/manual_en_v3.pdf [accessed 2020-10-19]
38. Tsuya A, Sugawara Y, Tanaka A, Narimatsu H. Do cancer patients tweet? Examining the twitter use of cancer patients in Japan. J Med Internet Res 2014 May 27;16(5):e137 [FREE Full text] [doi: 10.2196/jmir.3298] [Medline: 24867458]
39. Ministry of Health, Labour and Welfare. Overview of medical facility (static / dynamic) surveys and hospital reports 2017 [In Japanese]. 2017. URL: https://www.mhlw.go.jp/toukei/saikin/hw/iryosd/17/ [accessed 2020-10-19]
40. Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. Bone Marrow Transplant 2013 Mar;48(3):452-458 [FREE Full text] [doi: 10.1038/bmt.2012.244] [Medline: 23208313]
41. MSD K.K. "Health and Medical Awareness Survey" [In Japanese]. 2017. URL: http://www.msd.co.jp/static/pdf/corporate_20171127_2.pdf [accessed 2020-10-17]
42. National Research Corporation. 1 in 5 Americans Use Social Media for Health Care Information. 2011. URL: http://hcmg.nationalresearch.com/public/News.aspx?ID=9 [accessed 2019-01-08] [WebCite Cache ID 72AhcT7l7]
43. Grajales FJ, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 2014 Feb 11;16(2):e13 [FREE Full text] [doi: 10.2196/jmir.2912] [Medline: 24518354]
44. Haruna N, Ako K. Special Issue: Medical Care Connected with Social Media [In Japanese]. Nikkei Medical 2011;40(7):58-74.

45. Alexander DE. Social media in disaster risk reduction and crisis management. Sci Eng Ethics 2014 Sep;20(3):717-733. [doi: 10.1007/s11948-013-9502-z] [Medline: 24306994]
46. Huang C, Chan E, Hyder AA. Web 2.0 and internet social networking: a new tool for disaster management? Lessons from Taiwan. BMC Med Inform Decis Mak 2010 Oct 06;10:57 [FREE Full text] [doi: 10.1186/1472-6947-10-57] [Medline: 20925944]
47. Acar A, Muraki Y. Twitter for crisis communication: lessons learned from Japan's tsunami disaster. IJWBC 2011;7(3):392. [doi: 10.1504/ijwbc.2011.041206]
48. Vieweg S, Hughes A, Starbird K, Palen L. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. 2010 Presented at: CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April, 2010; Atlanta Georgia USA URL: https://doi.org/10.1145/1753326.1753486
49. Houston JB, Hawthorne J, Perreault MF, Park EH, Goldstein Hode M, Halliwell MR, et al. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. Disasters 2015 Jan 22;39(1):1-22. [doi: 10.1111/disa.12092] [Medline: 25243593]
50. Côté E, Hearn R. The medical response to the Boston Marathon bombings: an analysis of social media commentary and professional opinion. Perspect Public Health 2016 Nov;136(6):339-344. [doi: 10.1177/1757913916644480] [Medline: 27161388]
51. Cassa C, Chunara R, Mandl K, Brownstein JS. Twitter as a sentinel in emergency situations: lessons from the Boston marathon explosions. PLoS Curr 2013 Jul 02;5 [FREE Full text] [doi: 10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b] [Medline: 23852273]
52. Bekkat-Berkani R, Romano-Mazzotti L. Understanding the unique characteristics of seasonal influenza illness to improve vaccine uptake in the US. Vaccine 2018 Nov 19;36(48):7276-7285 [FREE Full text] [doi: 10.1016/j.vaccine.2018.10.027] [Medline: 30366802]
53. Kmenta M, Zetter R, Berger U, Bastl K. Pollen information consumption as an indicator of pollen allergy burden. Wien Klin Wochenschr 2016 Jan;128(1-2):59-67. [doi: 10.1007/s00508-015-0855-y] [Medline: 26373744]
54. Sugawara Y, Narimatsu H, Tsuya A, Tanaka A, Fukao A. Medical Institutions and Twitter: A Novel Tool for Public Communication in Japan. JMIR Public Health Surveill 2016;2(1):e19 [FREE Full text] [doi: 10.2196/publichealth.4831] [Medline: 27227154]
55. Greaves F, Laverty AA, Cano DR, Moilanen K, Pulman S, Darzi A, et al. Tweets about hospital quality: a mixed methods study. BMJ Qual Saf 2014 Oct;23(10):838-846 [FREE Full text] [doi: 10.1136/bmjqs-2014-002875] [Medline: 24748372]
56. 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]
57. Schumacher KR, Lee JM, Pasquali SK. Social media in paediatric heart disease: professional use and opportunities to improve cardiac care. Cardiol Young 2015 Dec;25(8):1584-1589. [doi: 10.1017/S1047951115002292] [Medline: 26675608]
58. Cain J. Social media in health care: the case for organizational policy and employee education. Am J Health Syst Pharm 2011 Jun 01;68(11):1036-1040. [doi: 10.2146/ajhp100589] [Medline: 21593233]
59. Gagnon K, Sabus C. Professionalism in a digital age: opportunities and considerations for using social media in health care. Phys Ther 2015 Mar;95(3):406-414. [doi: 10.2522/ptj.20130227] [Medline: 24903111]
60. Pillow MT, Hopson L, Bond M, Cabrera D, Patterson L, Pearson D, Council of Residency Directors Social Media Task Force. Social media guidelines and best practices: recommendations from the Council of Residency Directors Social Media Task Force. West J Emerg Med 2014 Feb;15(1):26-30 [FREE Full text] [doi: 10.5811/westjem.2013.7.14945] [Medline: 24578765]
61. British Medical Association. Using social media: practical and ethical guidance for doctors and medical students. 2011. URL: http://www.medschools.ac.uk/SiteCollectionDocuments/social_media_guidance_may2011.pdf [accessed 2015-11-24] [WebCite Cache ID 6dH18d3p1]
62. Ministry of Economy, Trade and Industry. METI Compiled a Survey Report on Business Activities Utilizing Social Media. 2016. URL: https://www.meti.go.jp/english/press/2016/0411_01.html [accessed 2020-03-21] [WebCite Cache ID 715h0JMgy]
63. Yoshihiro I, Yuto T. The management of SNS marketing strategy in Japan [In Japanese]. Bulletin of Yamagata University (Social Science) 2014;45(1):91-127.
64. Medical Institutions Internet Patrol. Ministry of Health, Labour and Welfare Commissioned Project Strengthening Monitoring System of Websites Related to Medical Service, Medical Institutions Internet Patrol [In Japanese]. URL: http://iryoukoukoku-patroll.com/ [accessed 2020-10-19]
65. Ministry of Health, Labour and Welfare. The Internet Patrol Project (2017) [In Japanese]. URL: https://www.mhlw.go.jp/file/05-Shingikai-10801000-Iseikyoku-Soumuka/0000209654.pdf [accessed 2020-10-19]
66. Omair A. Sample size estimation and sampling techniques for selecting a representative sample. J Health Spec 2014;2(4):142. [doi: 10.4103/1658-600x.142783]
67. Lombard M, Snyder-Duch J, Bracken CC. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. Human Comm Res 2002 Oct;28(4):587-604. [doi: 10.1111/j.1468-2958.2002.tb00826.x]
68. Blei DM. Probabilistic topic models. Commun. ACM 2012 Apr 01;55(4):77-84. [doi: 10.1145/2133806.2133826]

69. Mimno D, Wallach H, Talley E, Leenders M, McCallum A. Optimizing Semantic Coherence in Topic Models. 2011 Jul Presented at: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; July 27–31, 2011; Edinburgh, Scotland, UK.
70. Hanson C, West J, Neiger B, Thackeray R, Barnes M, McIntyre E. Use and Acceptance of Social Media Among Health Educators. American Journal of Health Education 2013 Jan 23;42(4):197-204. [doi: 10.1080/19325037.2011.10599188]
71. The Nikkei. Is "LINE" made in Japan? Made in Korea?. 2013. URL: https://www.nikkei.com/article/DGXNASFK2203C_S3A120C1000000/ [accessed 2017-10-03]
72. Ministry of Internal Affairs and Communications, Institute for Information and Communications Policy. Survey on usage time and information behavior of information and communication media 2016 [summary] [In Japanese]. 2017. URL: http://www.soumu.go.jp/main_content/000492876.pdf [accessed 2019-08-05]
73. The Nishinippon Shimbun. Iizuka Hospital distributes information about medical care, health and events on LINE every Wednesday [In Japanese]. 2017. URL: https://www.nishinippon.co.jp/nnp/medical/article/344738/ [accessed 2017-07-22]
74. Hu Y, Manikonda L, Kambhampati S. What We Instagram: A First Analysis of Instagram Photo Content and User Types. Palo Alto, California: The AAAI Press; 2014 Presented at: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media; June 1–4, 2014; Ann Arbor, Michigan, USA.
75. Martinez-Millana A, Fernandez-Llatas C, Basagoiti Bilbao I, Traver Salcedo M, Traver Salcedo V. Evaluating the Social Media Performance of Hospitals in Spain: A Longitudinal and Comparative Study. J Med Internet Res 2017 May 23;19(5):e181 [FREE Full text] [doi: 10.2196/jmir.6763] [Medline: 28536091]

Abbreviations

MAU: Monthly Active Users
MHLW: Ministry of Health, Labour and Welfare

Edited by G Eysenbach, Q Zeng; submitted 11.03.20; peer-reviewed by A Martinez-Millana, P Delir Haghighi, JR Bautista; comments to author 28.08.20; revised version received 21.10.20; accepted 25.10.20; published 27.11.20.

Please cite as:
Sugawara Y, Murakami M, Narimatsu H
Use of Social Media by Hospitals and Clinics in Japan: Descriptive Study
JMIR Med Inform 2020;8(11):e18666
URL: https://medinform.jmir.org/2020/11/e18666
doi:10.2196/18666
PMID:33245281

Original Paper

Automatic Structuring of Ontology Terms Based on Lexical Granularity and Machine Learning: Algorithm Development and Validation

Lingyun Luo[1,2*], PhD; Jingtao Feng[1*], BS; Huijun Yu[3], BS; Jiaolong Wang[1], BA

[1]School of Computer Science, University of South China, Hengyang, China

[2]Hunan Medical Big Data International Science and Technology Innovation Cooperation Base, Hengyang, China

[3]Clinical Laboratory Medicine Center, Shenzhen Hospital, Southern Medical University, Shenzhen, China

[*]these authors contributed equally

Corresponding Author:
Lingyun Luo, PhD
School of Computer Science
University of South China
28 West Changsheng Rd
Hengyang, 421001
China
Phone: 86 7348282473
Email: luoly@usc.edu.cn

Abstract

Background: As the manual creation and maintenance of biomedical ontologies are labor-intensive, automatic aids are desirable in the lifecycle of ontology development.

Objective: Provided with a set of concept names in the Foundational Model of Anatomy (FMA), we propose an innovative method for automatically generating the taxonomy and the partonomy structures among them, respectively.

Methods: Our approach comprises 2 main tasks: The first task is predicting the direct relation between 2 given concept names by utilizing word embedding methods and training 2 machine learning models, Convolutional Neural Networks (CNN) and Bidirectional Long Short-term Memory Networks (Bi-LSTM). The second task is the introduction of an original granularity-based method to identify the semantic structures among a group of given concept names by leveraging these trained models.

Results: Results show that both CNN and Bi-LSTM perform well on the first task, with F1 measures above 0.91. For the second task, our approach achieves an average F1 measure of 0.79 on 100 case studies in the FMA using Bi-LSTM, which outperforms the primitive pairwise-based method.

Conclusions: We have investigated an automatic way of predicting a hierarchical relationship between 2 concept names; based on this, we have further invented a methodology to structure a group of concept names automatically. This study is an initial investigation that will shed light on further work on the automatic creation and enrichment of biomedical ontologies.

(JMIR Med Inform 2020;8(11):e22333) doi:10.2196/22333

KEYWORDS

ontology; automatic structuring; Foundational Model of Anatomy; lexical granularity; machine learning

Introduction

Background

Biomedical ontologies are formalized representations of concepts and the relationships among these concepts for the biomedical domain, and they play a vital role in many medical settings [1]. The constructions of ontologies are labor-intensive and time-consuming. In addition, their evolvements often require concept enrichment that must be manually reviewed by domain experts. Thus, automatic mechanisms are desirable in both ontology construction and ontology maintenance tasks.

In recent years, many ontology learning (OL) efforts have been made to automate the construction of ontologies from free text [2]. An important subtask in the OL process is relation extraction that aims to extract a novel relationship between known concepts [3]. Putting aside the accuracy of extraction, the discovery of

semantic relations from text has its drawbacks: one is that the representations of concepts and relations in the text are usually nonstandard, and the other is that the knowledge extracted from text is often limited and not curated. Due to the widespread use of biomedical ontologies [4], their quality has become very important [5]. As such, in this study, instead of discovering semantic relations from extrinsic information, we investigate an automatic way of uncovering relations between ontology concept names by leveraging the intrinsic knowledge of the ontology itself.

An important observation of biomedical ontologies is that the lexical patterns of the concepts often indicate, to a certain degree, the structural relations between them, especially for hierarchical relations. For instance, in the Foundational Model of Anatomy (FMA) [6], *Left hemidiaphragm* is part of *Diaphragm*, and *Superior mediastinal lymph node* is a *Mediastinal lymph node*. We can notice that in each example, the parent concept name is a substring of the child concept name, as the parent is semantically more general than the child. Using naming conventions in biomedical ontologies is a principle recommended by the Open Biological and Biomedical Ontology (OBO) Foundry [7]. In the literature, lexical-structural relevance had been leveraged for many ontology-related tasks. For instance, we used subphrases of concept names and structural information for disambiguating terms in the FMA [8]. Also, the approach of combining lexical and structural methods is widely adopted in many ontology auditing studies [9-11]. Note that in this paper, we use the terms “concept name” and “term” interchangeably.

In this study, we propose an automatic approach for structuring a given set of concept names based on their lexical granularity. We started by investigating an automatic way to predict the direct relation between 2 given concepts by employing machine learning (ML) algorithms. Since word embedding tools such as Word2Vec [12] and Bert-as-service [13] can extract the semantic features of words and encode the words into feature vectors, relations between words are retained to some extent. By feeding encoded term pairs along with their corresponding relations into ML models such as Convolutional Neural Networks (CNN) [14], Long Short-term Memory Networks (LSTM) [15], or Support Vector Machine (SVM) [16], we can train the models as classifiers to predict the relations between given concept names.

We selected the most common hierarchical relations in biomedical ontologies for experiments: the *is-a* relation and the *part-of* relation. The training dataset comprised randomly selected pairs from the taxonomy and partonomy of the ontologies. Each pair was either directly related by *is-a* or by *part-of*. In addition, we added a third type of concept pairs to the training set: concept pairs that are not directly related (*ndr*). For each pair in the training set, we encoded the 2 terms to vectors using Bert-as-service [13] at first. The subtraction of the 2 vectors formed an input instance for ML models. After training, the models were able to classify a given term pair (A, B) into one of the 3 classes: (A *is-a* B), (A *part-of* B), or (A *ndr* B).

Moving forward, provided with a group of concept names, we aimed to determine how to structure them automatically by utilizing the above ML classifiers. Intuitively, the relative positions of all the concepts can be achieved by pairwise comparisons. However, pairwise comparisons will not only increase the algorithm complexity but also tend to introduce false-positive relations. To deal with this problem, we deployed our previous work [11] on concept granularity to obtain the positions of concepts: Firstly, we determined all the parallel concept sets (PCSs) in the given names. Secondly, we placed them into different hierarchical levels based on their granularity, forming PCS threads. Each thread determined a PCS hierarchy. Lastly, we used the above ML models to determine the relations between neighboring terms along the threads as well as relations between certain terms from different threads. As a result, we achieved the goal of predicting the whole taxonomy and partonomy structures for the given names. To the best of our knowledge, this is the first study that investigates automatic semantic structure generation for a group of concept names in biomedical ontologies.

Related Work

In the literature, automatic methods were proposed to alleviate human efforts from different aspects of the ontology lifecycle. Many researchers utilized automatic methods to facilitate semantic knowledge extraction for ontology enrichment. For example, Pembeci et al [17] proposed a supervised ontology enrichment algorithm by using concept similarity scores computed via Word2Vec models to discover other related concepts for a given concept. We refer to Liu et al [18] for more references. For ontology concept name prediction, Zheng et al [19] explored deep learning-based approaches to automatically suggest new concept names in the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT), under the condition that a bag of words is given. However, only a few studies worked on automating relation prediction and concept organization within ontologies. Zheng et al [20] verified whether an *is-a* link should exist between a new child concept and an existing parent concept in the SNOMED CT. Liu et al [21] proposed a CNN-based method to support the insertion of new concepts into the SNOMED CT: The CNN classifier was trained by vectors translated from concepts using the Doc2Vec algorithm. Afterward, it was able to decide if a given concept has the *is-a* relation with existing concepts in the ontology. Later, they also used a transfer learning method based on BERT to support the insertion of new concepts [22]. A limitation of the work is that at least one parent had to be given for the concept to be inserted beforehand.

Our study differs from the above work mainly in the following aspects: (1) Instead of predicting the insertion place of a new concept or predicting the relation between a particular concept pair, we predict the whole hierarchical structure for a given set of concept names; (2) aside from names of the concepts, we do not need extra information to predict their positions in the whole group; and (3) instead of concatenating the child and the parent, we encode them separately and use their subtraction as an input instance for the ML models.

Methods

Materials

We tested our methodology in the FMA [6], which is both a theory of human anatomy and an ontology artifact. In particular, it is a representation of the canonical, phenotypic structure of the human body and its typical components at all biological levels. It is a model suitable for machine manipulation with more than 100,000 concepts, including macroscopic, microscopic, and subcellular canonical anatomy.

For our analysis, we used version 5.0.0 of the FMA (Structural Informatics Group at the University of Washington) [23]. It is distributed as Web Ontology Language (OWL) files, which enables the FMA to be stored in resource-description-frame (RDF) data stores and made available for querying via SPARQL [24]. In this study, we used Virtuoso (version 7.2.5.1; OpenLink Software) as our RDF store [25].

Model Training and Testing for Direct Relation Prediction

Data Preparation

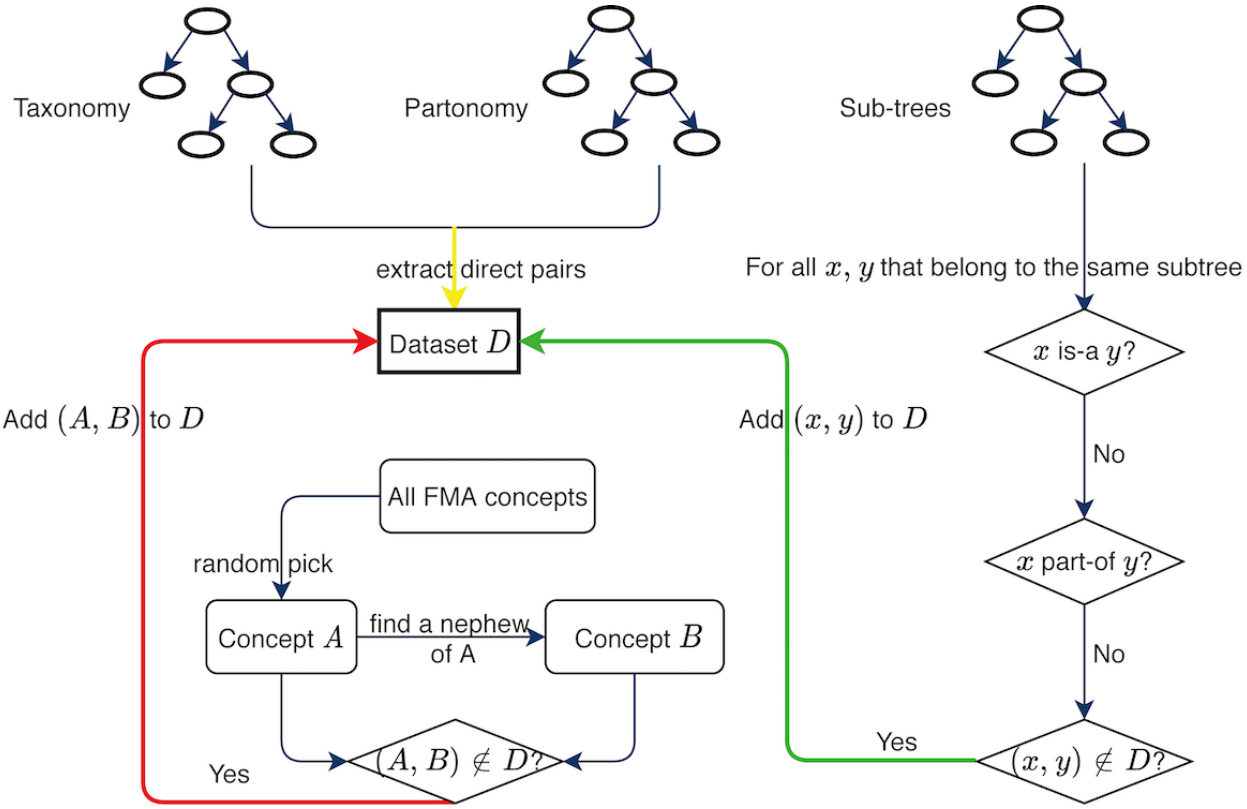
We use the FMA to describe the data preparation process without a loss of generality. We first extracted all the concept pairs directly related by *is-a* or *part-of* from the FMA. The resulting set, *D*, contained 104,665 *is-a* pairs and 61,878 *part-of* pairs. All the children from *D* comprised a set *C*, and all the parents from *D* comprised a set *P*. We then generated the third type of pairs, which were pairs that are not directly related (*ndr*).

The *ndr* pairs consisted of 2 kinds: (1) pairs of terms that share the same ancestor, and (2) pairs comprising a random concept A in the FMA and a child of the sibling of A (ie, uncle-nephew pairs). For the first kind of pair (pairs of terms that share the same ancestor), we first found all the subtrees in the FMA with sizes between 60 and 135. Then, for each of these trees, we let all of its node terms pair with each other. If a direct *is-a* or *part-of* relation did not connect the 2 elements of each pair, it was an *ndr* pair added to the dataset *D*.

The reason that we chose these 2 kinds of *ndr* pairs are the following: Since our ultimate goal was to organize a group of closely related terms, *ndr* pairs in the training dataset should not just be chosen at random. Thus, we intentionally included *ndr* pairs that originated from the same subtrees into the dataset, as the first kind of *ndr* pairs do, to help recognize *ndr* relations in the target groups. As the subtrees should be neither too large nor too small, only subtrees with moderate sizes between 60 and 135 were selected for our experiment. Note that although *is-a* and *part-of* are both transitive relations, indirect *is-a* pairs and indirect *part-of* pairs were classified as *ndr* pairs. For the second kind of *ndr* pairs, we included certain uncle-nephew pairs from the whole FMA dataset, as they tend to be mispredicted to have parent-child relations.

The data preparation process is illustrated in Figure 1. Our selection process of *ndr* pairs stopped when the number of *ndr* pairs reached 3 times the summation of the numbers of *is-a* pairs and *part-of* pairs. The ratio of these 2 kinds of *ndr* pairs was 1:1; that is, we randomly selected 249,815 pairs from the first kind of *ndr* pairs and the same number from millions of uncle-nephew pairs. The number of *ndr* pairs was set much larger than the numbers of *is-a* pairs and *part-of* pairs to better match the real situations in the ontology.

Figure 1. The data preparation process. The final dataset D consists of 3 parts: (1) all of the direct *is-a* and *part-of* pairs in the Foundational Model of Anatomy (inputted by the yellow arrow); (2) *ndr* pairs of terms that share the same ancestor (inputted by the green arrow); and (3) *ndr* uncle-nephew pairs (inputted by the red arrow).



Embedding

Our aim was to train an ML algorithm that was able to determine whether 2 given ordered terms maintain 1 of the 3 relations between them, namely, an *is-a* relation, a *part-of* relation, or an *ndr* relation. Above all, the term pairs needed to be converted to vectors, which is called embedding. To do this, firstly, we used the Bert-as-service tool [13] to acquire the vector representations for all words that appeared in the dataset *D*. Each word was represented by a 768-length vector; thus, each concept name in *D* was represented by a sequence of vectors. Secondly, to align all the concept names, we padded all the sequences' vectors to the same length of 20. Lastly, all the child vectors were subtracted from their respective parent vectors to create the input vectors for classification algorithms. We selected subtraction rather than concatenation because subtraction would catch the differentiation between the parent and the child. As a result, each input vector took shape (1,20,768) and was labeled by its corresponding relation.

Model Training and Direct Relation Prediction

We shuffled the input vectors along with their labels and used 80% of them as the training set, 10% of them as the validation set, and the remaining 10% as the testing set. Since FMA terms are all short texts, we selected the classic TextCNN proposed by Yoon Kim [26], which is widely used in short-text classification like our CNN model. The other classification model we used was Bidirectional Long Short-term Memory Networks (Bi-LSTM) [15], which is often used to model contextual information in natural language processing tasks. In our experiments, the parent term and the child term were used as contextual information for Bi-LSTM to predict the relationship between them.

We ran the models using Keras [27] on CentOS with 240 GB of memory and 4 Tesla M60.

In the CNN model, we used 3 Conv1D layers and 2 MaxPooling1D layers following the input layer. After flattening the last layer's output, we added 2 dense layers such that the former had a *relu* activation and the latter had a *softmax* activation. The cost function we leveraged was *categorical-crossentropy* in Keras. After training, for each input vector that represents a pair of concept names, the CNN model would predict a relation between the 2 concepts.

The second classification model we used was Bi-LSTM. After the input layer, we added a Bi-LSTM layer with 32 memory units in the middle. Then, we flattened the output of the last layer to add a dense layer which had a *softmax* activation. Cost function *categorical- crossentropy* in Keras was also used for classification.

For both models, we set the training data to batches of size 512 and set the *epoch* parameter as 50. For each iteration, we used the validation data to evaluate the model's performance.

The testing set was used to evaluate the performance of each model. By comparing the predicted results with the real situations in the FMA, we calculated metrics such as the precision, recall, and F1 scores for each model separately.

To demonstrate the robustness of our trained models, we repeated the above experiment 100 times and obtained the average precision, recall, and F1 values. The training set, validation set, and testing set were randomly divided each time, but the 8:1:1 ratio was maintained. In the following step, we selected a particular group of terms from each testing set and automatically obtained the taxonomy and partonomy structures among those terms.

Automatic Structuring for a Group of Concept Names

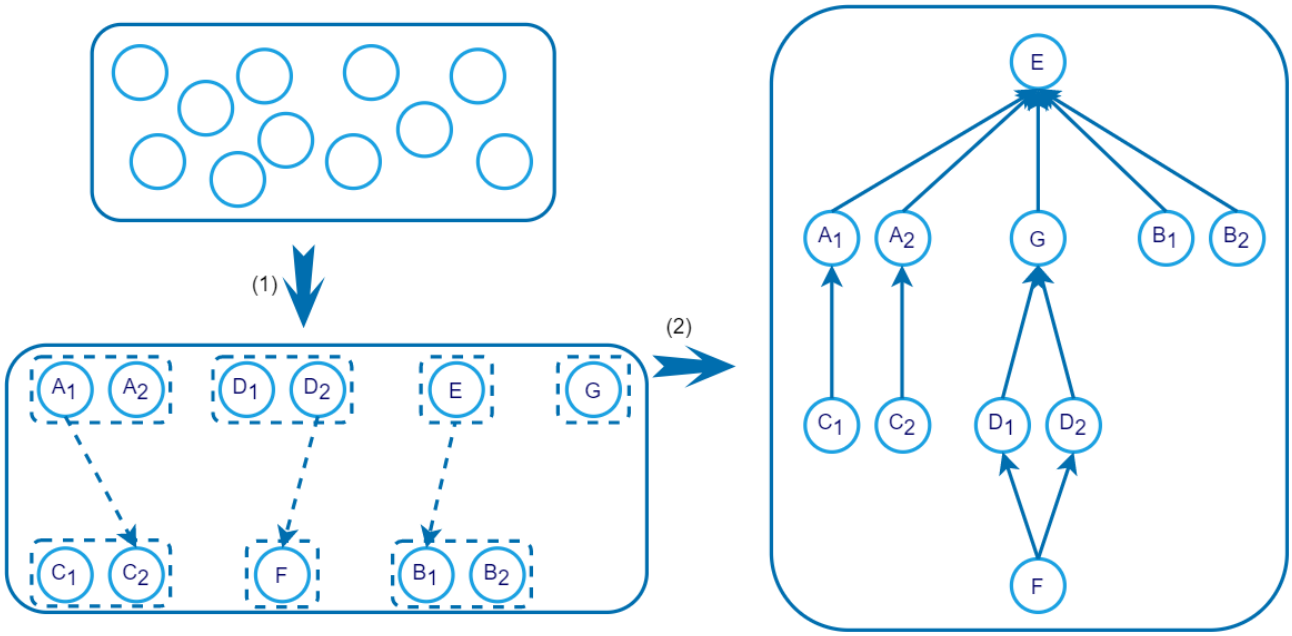
Algorithm Overview

The above ML models only predict if 2 given terms are directly related by *is-a* or *part-of*. However, rather than predicting the relation between 2 random terms, a more meaningful use lies in the organization of a given set of closely related terms. To achieve this goal, we needed to obtain their relative positions.

An intuitive solution is to use the pairwise comparison. Let the target term set be Q. Suppose the number of terms in Q is *M*; we will need $M(M-1)$ times of testing to obtain the pairwise relations among them. However, apart from time complexity, another problem with this solution is that it will introduce too many *ndr* pairs since real *is-a* or *part-of* relations in Q are quite sparse. Thus, the prediction results for *ndr* pairs will easily affect the prediction results for *is-a* and *part-of* pairs.

As such, to reduce the use of pairwise comparison, we deployed our previous work [11] on concept granularity to obtain the relative positions of the terms. Specifically, we divided the target set into small parallel concept sets (PCSs) [11]. As parallel concepts remain at the same level of granularity, it turned out that, in the end, we only needed to organize the PCSs. To do this, we first placed the PCSs into different hierarchical levels based on their granularity, forming PCS threads. Each thread determined a semantic hierarchy. Then, we determined the hierarchy between different threads. Then, after the whole structure was obtained, we utilized the trained ML models to predict the relations between directly connected terms and thus obtained the whole semantic map. This procedure is briefly illustrated in Figure 2.

Figure 2. The use of lexical granularity to obtain the relative positions of terms. (1) Parallel concept sets (PCS) and PCS thread detection; 7 PCS nodes and 4 PCS threads were detected in this example. PCS: represented by dashed rectangles; Concept names: represented by circles; Substring relations: represented by dashed arrows. (2) Relation prediction. *is-a* or *part-of* relations predicted by the classification model: represented by solid arrows.



PCS Detection

A parallel concept set (PCS) is a set comprised of concepts sharing the same level of conceptual knowledge [11], such as symmetric concepts. A pair of concepts is called symmetric if the concept names are the same but for the possible difference in a single occurrence of the modifiers used [10]. For instance, *Lower extremity part-Upper extremity part* is a symmetric concept pair concerning the symmetric modifier pair *Upper* and *Lower*.

In order to detect all the symmetric concept pairs in *Q*, we needed to retrieve all the symmetric modifier pairs first. To do this, we used the Stanford Parser [28] to obtain all the noun-phrase (NP) chunks without prepositions. For all the modifiers in those chunks, any 2 of them that share a common context were selected to form a modifier pair. After retrieving all the symmetric modifier pairs from *Q*, we easily detected all the symmetric concepts using SPARQL queries [24]. In the end, every symmetric term pair formed a PCS. For terms whose symmetric counterparts could not be found in *Q*, each of these formed a PCS by itself.

PCS Thread Detection

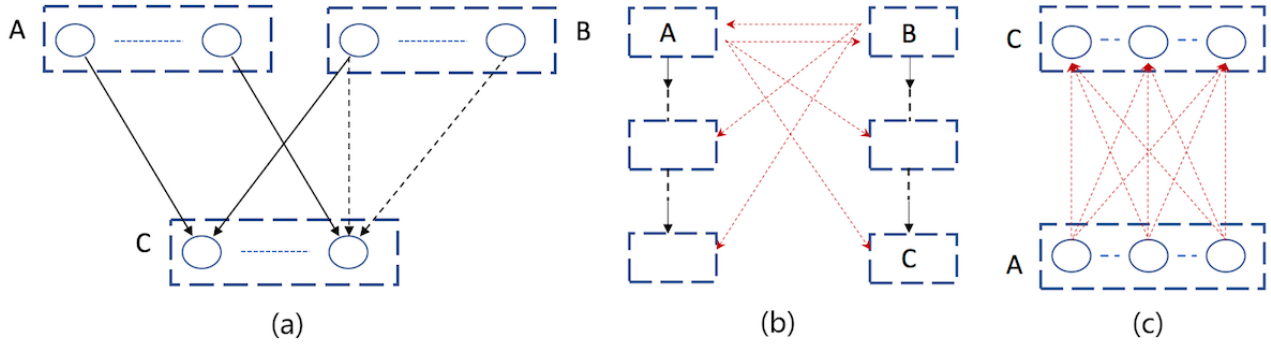
As noted, for hierarchical relations, the parent term is more general than the child term and is usually a substring of the child term. As a result, we can leverage the substring threads in Q to organize the PCSs identified from the above step. We used $A \sqsubset B$ to represent that *A* is a substring of *B*. A substring thread $A_0 \sqsubset A_1 \ldots \sqsubset \ldots A_{n-1} \sqsubset A_n$ would correspond to a parent-child

thread A_0 A_1... A_{n-1} A_n. Along each parent-child thread, we generalized every term node to the PCS that the term belonged to. As a result, all the PCSs were organized into several threads. Each thread was named a PCS thread. Note that some threads may only contain 1 PCS node. Also, some PCSs may appear in several threads.

Relation Prediction

The relative positions of nonroot PCS nodes were determined. Hence, we no longer looked for their parents elsewhere but only predicted relations between concepts in neighboring nodes. Specifically, we first paired each term in the PCS with its substring term in the parent PCS. If no substring term was found in the parent PCS, the term would be paired with every item in the parent. As illustrated in Figure 3a, A, C, and B, C were neighboring nodes along 2 PCS threads, respectively. Since the right-most term in C had no substring in B, it was paired with every term in B, represented by dashed arrows. Then, we predicted the relations between the paired terms by leveraging the previously trained classification models.

Figure 3. Determination of term pairs to be fed into machine learning models for relation prediction. Parallel concept sets (PCSs): represented by rectangles; Concept names: represented by ovals. (a) A, B, and C are 3 PCS nodes; A, C and B, C are neighboring nodes along 2 PCS threads, respectively. Substring relations: represented by solid arrows. As the right-most term C has no substring term in B, it is paired with every term in B, represented by dashed arrows. Each arrow (solid or dashed) connects 2 terms such that the relation between them is predicted using classification models. (b) A and B are 2 different PCS thread roots. Each root is paired with every PCS node in other threads under different roots; red dashed arrows are used to connect them. For instance, (C, A) is such a pair. (c) Classification models are used to predict the pairwise relations between concept names in C and A from the above step.



In regard to the PCS thread roots, if all the threads shared 1 root, no further treatment was needed. If there existed more than 1 different thread root, we still leveraged pairwise comparison to determine the parents for the roots: For each PCS thread root, we first paired it with every PCS node in other threads, as illustrated in Figure 3b. For instance, (C, A) was such a pair, with C as the parent PCS and A as the child PCS (Figure 3c). We used the ML models to predict the pairwise relations between terms in those paired PCS nodes. As Figure 3c illustrates, each term in C was paired with every term in A, and the specific relation between each pair would be predicted by the previously trained classification models.

Lastly, only *is-a* and *part-of* edges would be retained. For the given group of terms, suppose the number of predicted *is-a* relations was *P*, and the number of correct ones among them was *CP*; then, the precision for *is-a* was calculated as *CP/P*. Further, suppose the number of original *is-a* relations in the FMA was *O*, and the number of them that were correctly predicted was *CO*; then, the recall for *is-a* was *CO/O*.

Case Studies

To test the generalizability of our method, we selected a group of terms from each of the testing sets in the 100 cross-validation experiments for automatic structuring. As mentioned, the most useful scenario happens when the terms are closely related instead of semantically distant. Thus, we only selected terms that belong to the same tree for experiments.

The process was as follows: Firstly, we collected all the term roots in the testing set and collected all the *is-a* and *part-of* descendants under them, forming a concept tree for each root. Secondly, we picked out the trees with more than 20 elements. Lastly, we randomly selected a tree and created a set formed by all the terms in that tree as our study case. Note that we manually assured that none of the concepts in the case studies had appeared in the training set.

For the selected 100 cases, we followed the steps described above to predict the whole semantic map among the concept names in the groups. Our experiments separately leveraged the 2 previously trained models for direct relation prediction. By comparing the predicted results with the real cases in the FMA, we evaluated the performance of our methodology by calculating the average precision, recall, and F1 values for all the cases.

For a more specific analysis of the results, we selected the largest case with root “First Rib” among the 100 cases. The set contained 57 concepts with 89 relations among them in the FMA, including 34 *is-a* relations and 55 *part-of* relations, as shown in Multimedia Appendix 1.

To demonstrate the advantage of our PCS-based method, we performed another group of experiments for the case study on “First Rib” based on primitive pairwise comparisons among the whole set of concept names. We fed 3192 (from 57×56) term pairs to the models for direct relation predictions. Then, a comparison between the PCS thread-based method and the primitive pairwise-based method was made for this case.

Results

Classifiers Can Predict the Direct Relation Between 2 Given Concept Names

Using the remaining 10% of the data as the testing set in each of the cross-validation experiments, we evaluated the performances of the 2 models on direct relation prediction between 2 given concept names. The average results are shown in Table 1. The table shows that the models performed well on this task, with both precision and recall above 0.9. This demonstrates that machine learning models, when trained by existing relations in the ontology, can be very effective at predicting relations between new incoming concepts, provided that their names are given. Based on this result, we invented the PCS thread-based method and further investigated the possibility of organizing a group of terms.

Table 1. Average performances of the 2 models on direct relation prediction (100 rounds).

Model	*is-a*		*part-of*		*ndr*		Overall		
	P[a]	R[b]	P	R	P	R	P	R	F1
Bi-LSTM[c]	0.93	0.91	0.90	0.91	0.97	0.93	0.95	0.92	0.93
CNN[d]	0.91	0.90	0.89	0.90	0.94	0.92	0.92	0.91	0.91

[a]P: precision.

[b]R: recall.

[c]Bi-LSTM: Bidirectional Long Short-term Memory Networks.

[d]CNN: Convolutional Neural Networks.

Automatic Structuring of Groups of Closely Related Terms

In the 100 testing sets, we found that the sizes of all trees were less than 60, and the 100 term groups we selected had an average size of 25. The smallest group contained 20 terms and the largest group contained 57 terms.

We applied the PCS-based algorithm to the 100 cases and calculated the average precision, recall, and F1 values for *is-a* and *part-of* based on Bi-LSTM and CNN, respectively. The results are shown in Table 2; the overall F1 score using Bi-LSTM was 0.79, which slightly outperformed the algorithm using CNN.

Table 2. Average performances of the parallel concept set (PCS) thread-based algorithm on 100-term groups.

Model	*is-a*		*part-of*		Overall		
	P[a]	R[b]	P	R	P	R	F1
Bi-LSTM[c]	0.84	0.79	0.82	0.68	0.83	0.76	0.79
CNN[d]	0.72	0.79	0.72	0.69	0.72	0.76	0.74

[a]P: precision.

[b]R: recall.

[c]Bi-LSTM: Bidirectional Long Short-term Memory Networks.

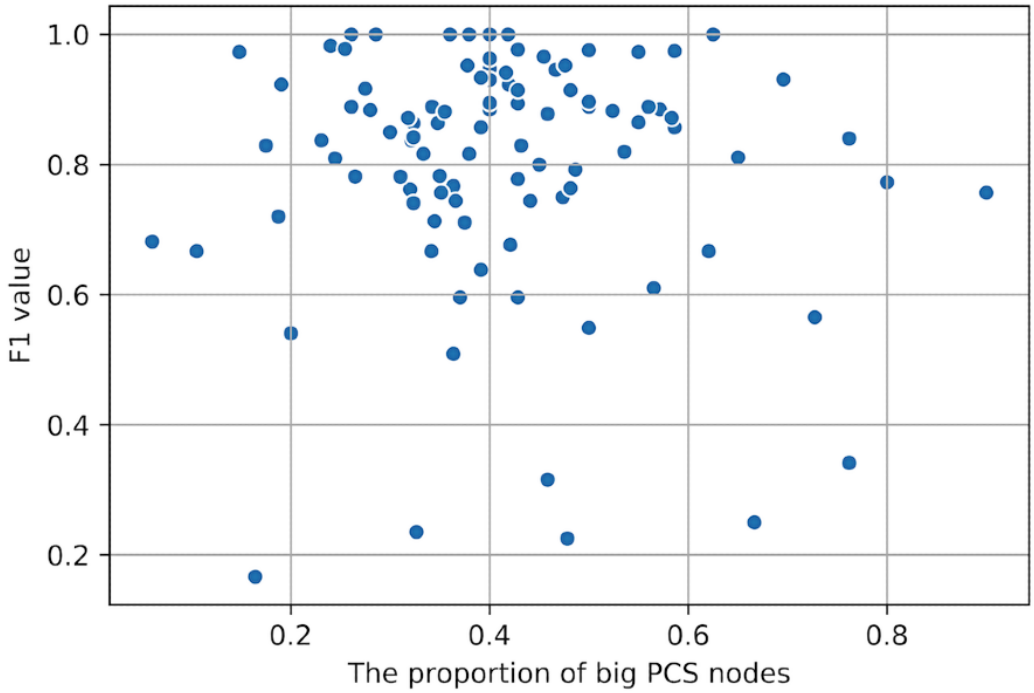
[d]CNN: Convolutional Neural Networks.

To analyze the influence of PCS nodes that contain at least 2 symmetric terms (ie, big PCS nodes) on the performances of the above algorithm, we calculated the proportion of big PCSs among all the PCSs for each study case and demonstrated the relation between the proportion and the F1 value (Figure 4). As it indicates, the PCS-based algorithm's performance does not have evident relevance with the richness of big PCS nodes.

Further, to demonstrate the usefulness of ML models in our approach, we collected all the *is-a* and *part-of* pairs without substring relationships in the 100 study cases and found 652 such pairs. Among the 652 pairs, 235 (36%) could be correctly predicted by our algorithm using both models. For instance, we correctly predicted the relation (*Endplate of intervertebral disk,is-a, Organ component)* in which the 2 terms have no shared word. In fact, the above ratio could have been much higher if the pairs had actually fed into the ML models, as some pairs without substring relationships were filtered out by the algorithm beforehand. On the other hand, the 100 cases contained 1140 *ndr* pairs in which 1 term is a substring of the other. Of the 1140 *ndr* pairs, 931 (82%) were correctly predicted by both models as *ndr* pairs.

As the above results show, our proposed algorithm works well on both term pairs, with or without obvious lexical patterns.

Figure 4. The relation between the proportion of big parallel concept set (PCS) nodes and the F1 value for 100 cases.



Specific Case Study on "First Rib"

For the specific case on "First Rib," in the 57 concept names to be structured, we detected 1 symmetric modifier pair, (*left*, *right*). We first divided all the concepts into 37 PCSs. Of these 37 PCSs, 20 PCSs contained 2 terms and 17 PCSs contained only 1 term. Then, based on lexical granularity, we found 29 PCS threads. Except for 1 thread that took "Fossa for first costal cartilage" as its root, all the other 28 threads shared the same root: "First Rib."

The results of the automatic structuring of term groups based on PCS threads and pairwise comparisons are shown in Table 3. The PCS thread-based method has higher precision, and the pairwise-based method has higher recalls. The reason is that the PCS thread-based method had filtered out certain pairs, including those with real *is-a* or *part-of* relations between their elements. On the other hand, the introduction of false-positive results was expected of the pairwise method.

We analyzed the results from the "First Rib" case for our PCS thread-based algorithm concerning the Bi-LSTM model to demonstrate why some relations were wrongly predicted or missed.

Table 3. The parallel concept set (PCS) thread-based algorithm versus the primitive pairwise-based algorithm on the "First Rib" case, using different models.

Model and Algorithm	*is-a*		*part-of*		Overall		
	Precision	Recall	Precision	Recall	Precision	Recall	F1
Bi-LSTM[a]							
Alg_1[b]	1.0	1.0	0.71	0.63	0.83	0.78	0.80
Alg_2[c]	0.94	1.0	0.55	0.90	0.66	0.94	0.78
CNN[d]							
Alg_1	0.97	1.0	0.73	0.64	0.83	0.78	0.80
Alg_2	0.58	1.0	0.42	0.98	0.47	0.98	0.64

[a]Bi-LSTM: Bidirectional Long Short-term Memory Networks.

[b]Alg_1: PCS thread-based algorithm.

[c]Alg_2: pairwise-based algorithm.

[d]CNN: Convolutional Neural Networks.

Result Analysis on "First Rib" for the PCS Thread-based Algorithm Concerning The Bi-LSTM Model

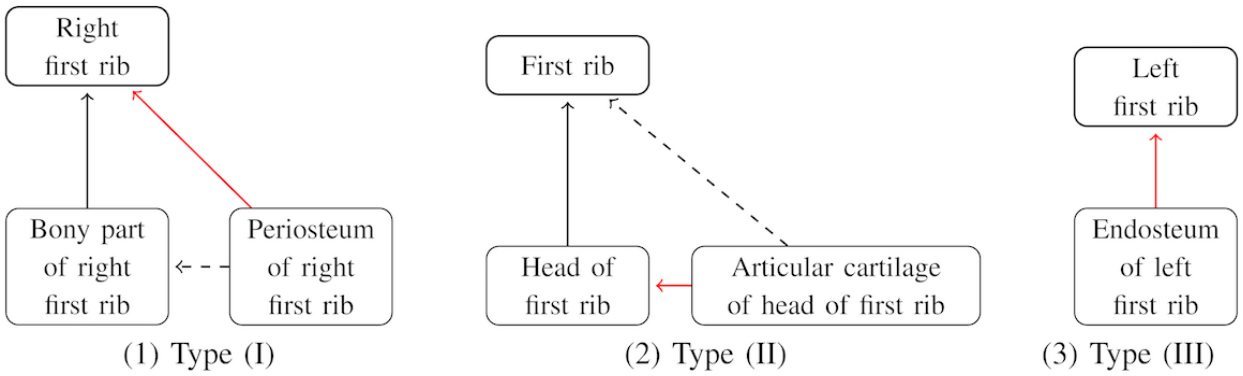
Using the Bi-LSTM model, our approach predicted 83 relations among the group of concept names, including 34 *is-a* relations and 49 *part-of* relations. The results are illustrated in Multimedia Appendix 2. All of the 34 *is-a* relations in the FMA were successfully discovered by the model.

Compared to the 55 real *part-of* relations in the FMA, 35 *part-of* relations were correctly predicted, which means that 20 *part-of* relations in the FMA were missed by our method and 14 predicted *part-of* relations were unexpected. As a result, we achieved an overall precision of 0.83 [from (34+35)/83] and a recall of 0.78 [from (34+35)/89], as shown in Table 3.

The 14 unexpected *part-of* relations that do not exist in the FMA can be divided into 3 types: (1) detected *part-of* relations that connected the child to a further parent than that of the FMA; (2) detected *part-of* relations that connected the child to a closer parent than that of the FMA; (3) detected *part-of* relations that did not exist in the FMA and had no counterpart relations in the FMA.

The first type was detected *part-of* relations that connected the child to a further parent than that of the FMA. As illustrated in Figure 5, Type I, *Periosteum of right first rib* has a closer parent, *Bony part of right first rib*, in the FMA, but the algorithm connected it to its ancestor, *Right first rib*. The reason is that the node *Bony part of first rib* is not lexically a substring of *Periosteum of right first rib* and thus did not appear in the corresponding PCS thread. Six predicted relations took this type.

Figure 5. Types of unexpected *part-of* relations that do not exist in the Foundational Model of Anatomy (FMA). Black solid arrows represent relations in the FMA successfully predicted by the model. Red arrows represent relations predicted by the model but not in FMA. Dashed arrows represent relations in FMA that were missed by the model. (1) Example of the first type of predicted relations; (2) example of the second type of predicted relations; (3) example of the third type of predicted relations.



The second type was detected *part-of* relations that connected the child to a closer parent than that of the FMA. As illustrated in Figure 5, Type II, the algorithm predicted the parent of *Articular cartilage of head of first rib* to be *Head of first rib*" instead of *First rib*. The reason is that only relations between neighboring terms along PCS threads would be predicted, but *Head of first rib* is in the middle of the other 2 terms. Six predicted relations took this type.

The third type was detected *part-of* relations that did not exist in the FMA and had no counterpart relations in the FMA. However, the parent term is a substring of the child term, as illustrated by the example in Figure 5, Type III. Two predicted relations took this type.

Although the above instances do not exist in the FMA, they are not all semantically wrong. For example, instances of the first type can be inferred from relation transitivity. Moreover, compared to the real cases in the FMA, the 6 instances of the second type were more reasonable because they show a finer granularity than their counterparts in the FMA. Also, the 2 instances of the third type were semantically correct.

On the other hand, the 20 missed *part-of* relations happened due to 1 reason: their parent-child term pairs were not fed to the model for prediction. As already described, for terms in nonroot PCS nodes, we only searched for their parents in neighboring parent PCSs. For the 20 missed cases, the parent and the child were not in neighboring PCSs and thus could not be discovered by our algorithm. For instance, in Figure 5, Type I, *Bony part of right first rib* and *Periosteum of right first rib* were not in neighboring PCS nodes along any thread, and thus, the pair was not fed to the model for relation prediction. Amongst the 20 missed cases, 11 cases came along with the instances of the first and second types of unexpected *part-of* pairs that did not exist in the FMA. As illustrated in Figure 5, Types I and II, while the model predicted an extra new relation, it would miss an old relation (a red arrow co-occurred with a dashed arrow). Only the first type of instance did not appear in the missed case because the intermediate parent was not in the term group.

If those missed parent-child term pairs were fed into the Bi-LSTM model, could they be correctly detected? Table 3 shows that the recall for Bi-LSTM was 0.94, which means that most of the original relations in the FMA could be successfully detected by the model if fed for prediction. However, as seen in the results, the precision values would drop greatly for primitive pairwise comparisons.

If the group of concept names to be structured do not show much relevance in their linguistic features, the number of PCS thread roots will increase. Under that circumstance, as our

algorithm pairs each root with every term in the other threads (Figure 3b), the number of term pairs fed to the ML models for relation prediction will increase. In the extreme case, all of the terms are roots by themselves, and the algorithm will turn into a pure pairwise-based algorithm. Fortunately, biomedical ontologies follow certain naming conventions, and meaningful usage of our methodology lies in the construction of a group of terms that are semantically close to each other; however, PCSs will play an important role in most cases.

Discussion

Principal Findings

This study proposes an innovative approach to the automatic construction of a given set of concept names with regard to *is-a* and *part-of* relations, which can save significant labor for domain experts in the ontology construction process. Our method comprises 2 main steps: (1) automatic prediction of direct semantic relation between 2 concept names using classification models; experiments on the FMA show that machine learning models can predict if 2 new terms are directly related by a *is-a* or *part-of* relation, provided that they are trained by existing relations; and (2) automatic construction of a group of closely related concept names based on PCS threads. First, we detected all the PCSs in the group and organized them into PCS threads based on lexical granularity. Second, we obtained the relative positions of different threads and the whole structure of the group. Lastly, we determined whether there exists an *is-a* relation or a *part-of* relation between each directly connected term pairs, thus completing the construction of the taxonomy and the partonomy structures.

Some concepts may have multiple *is-a* or *part-of* parents. As analyzed, for terms in nonroot PCS nodes, except for the threads they belong to, we do not look for their parents in other threads anymore. However, since PCS threads may have convergences, some terms may still be predicted to have multiple parents. In fact, no matter whether it is for *is-a* or *part-of*, parents that are not substrings of nonroot PCS nodes will be overlooked by our algorithm, such as the missed *part-of* instances in the FMA. On the other hand, all of the *is-a* relations were successfully discovered because all of the *is-a* parents appeared above their children along certain threads.

It is not a simple transition from step 1 to step 2. As shown by our results, even though the performances of ML models on relation prediction for randomly selected pairs may be quite promising (Table 2), it was still difficult to obtain the semantic structure for a set of terms using pure pairwise comparisons (Table 3). The reason is that pairwise comparisons introduce too many pairs: N terms will generate $N(N-1)$ pairs since direction matters. As the real connections among those terms can be quite sparse, most of the pairs are actually *ndr* pairs, which tremendously exceeds *is-a* pairs and *part-of* pairs. As such, even a small portion of *ndr* relations that were wrongly predicted as *is-a* or *part-of* relations could greatly decrease the precision of the results for *is-a* and *part-of* relations. That is why we introduced the PCS-based method, which only tested pairs that have a high possibility of exhibiting *is-a* or *part-of* relations between their elements. As a result, the number of false *is-a* relations and false *part-of* relations was reduced. However, on the other hand, the reduction of term pairs in the PCS-based stage has its drawback. As the step filters out many *ndr* pairs, it also misses some real relations between terms that are not lexically related, which is why the recall values for the PCS-based method were lower than that of the pairwise-based method.

Future Work

To improve the performance of our PCS-based method, we need to include more possible pairs to be inputted into the ML models for relation prediction. This requires a mechanism to be able to identify hierarchical relations between terms that are not lexically related, and in the meantime, to avoid introducing false-positive results. The difficulty lies in the ability to distinguish the *ndr* pairs from the other 2 relations. Future research may focus on the following aspects: (1) Although we enlarged the set of *ndr* pairs in this study, it is still impractical to collect all possible *ndr* pairs for classification; we will try ML algorithms that are able to classify the relations based only on positive samples, and hence, there will be no need to collect *ndr* pairs then. (2) Except for the lexical information of the concept names, we will try including additional knowledge such as metadata or even structural information to the embedding framework. (3) We tried 2 classic ML models in this study and did not apply too much effort to parameter tuning or model refinements. We believe that further exploration of this aspect will also help.

Also, to make the methodology provided in this study scalable to more cases in diverse ontologies such as SNOMED CT [29], the key is for the ML models to be able to “interpret” the semantic meaning behind each biomedical term. This will require a suitable embedding method in the biomedical field. In the future, we will try other embedding methods learned from multiple sources of biomedical data, such as Cui2Vec [30] or BioBert [31], to generalize the method to other cases.

The 100 cases we experimented with in the FMA are not large because the closely related term trees in the testing sets are relatively small. If the target group is much larger, the performance of the proposed algorithm may not be as strong since more terms will increase the number of *ndr* pairs. In the future, we will try the methods mentioned above and will work on larger term groups.

This study is an initial step toward automated ontology construction. As the training dataset is collected from the same ontology, the methodology we proposed in this study is applicable, provided that a part of the ontology is already known. To structure an ontology from scratch, the relations between entities will have to be learned from other knowledge sources such as the UMLS [32] or the literature. We believe our study will provide insight for future studies in this field. Moreover, the methodology provided here can be easily deployed for determining insertion positions for incoming concepts in ontology enrichment processes. Also, as the results show, some predicted relations are more reasonable than the real cases, which indicates that ontology quality assurance tasks can also benefit from this study.

Conclusions

In this study, given a set of closely related concept names in the FMA, we investigated an automatic way to generate the taxonomy and partonomy structures for them. We trained machine learning models to predict if there exists a direct hierarchical relation between 2 given terms; based on this, we further proposed an innovative granularity-based method to automatically organize a given set of terms. The 100 cases that we studied in the FMA demonstrated that our method is effective for structuring ontology concepts automatically, provided that their names are given. We believe this pioneering study will shed light on future studies on automatic ontology creation and ontology maintenance.

Acknowledgments

This work was supported in part by the Hunan Provincial Natural Science Foundation of China (No.2019JJ50520), the National Science Foundation of China (No.61502221), and the double first-class construct program of the University of South China (No.2017SYL16).

Authors' Contributions

LL conceived the idea and designed the algorithm. JF performed the experiments. LL and JF created the first manuscript. HY and JW contributed and revised later versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The original hierarchy map for the concept group “First Rib” in the Foundational Model of Anatomy. Red arrows represent 34 *is-a* relations; gray arrows represent 55 *part-of* relations.

[PNG File , 753 KB - medinform_v8i11e22333_app1.png]

Multimedia Appendix 2

Result for the automatic structuring of 57 concept names. The nodes represent the 37 parallel concept sets (PCSs); dashed rectangles represent PCSs with more than 1 term. The nodes in green are the 2 thread roots; arrows connect terms instead of PCSs. Gray arrows represent the 69 correctly predicted *is-a* and *part-of* relations; yellow arrows represent the 20 missed *part-of* relations; red arrows represent the 14 predicted *part-of* relations that do not exist in the Foundational Model of Anatomy (FMA).

[PDF File (Adobe PDF File), 35 KB - medinform_v8i11e22333_app2.pdf]

References

1. Kimura J, Shibasaki H. Recent Advances in Clinical Neurophysiology. In: Proceedings of the 10th International Congress of Emg and Clinical Neurophysiology. New York: Elsevier; 1995 Presented at: The 10th International Congress of EMG and Clinical Neurophysiology; October p. 15-19.
2. Al-Aswadi F, Chan H, Gan K. Automatic ontology construction from text: a review from shallow to deep learning trend. Artif Intell Rev 2019 Nov 08;53(6):3901-3928 [FREE Full text] [doi: 10.1007/s10462-019-09782-9]
3. Buitelaar P, Cimiano P, Magnini B. Ontology learning from text: methods, evaluation and applications. In: Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press; Jul 2005.
4. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform 2008:67-79 [FREE Full text] [Medline: 18660879]
5. Bodenreider O. Quality Assurance in Biomedical Terminologies and Ontologies. 2010 Apr 8 Presented at: A report to the Board of Scientific Counselors; Apr 2010; Bethesda URL: https://morc1.nlm.nih.gov/pubs/pres/20100408-BoSC-QA.pdf
6. Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003 Dec;36(6):478-500 [FREE Full text] [doi: 10.1016/j.jbi.2003.11.007] [Medline: 14759820]
7. Schober D, Smith B, Lewis SE, Kusnierczyk W, Lomax J, Mungall C, et al. Survey-based naming conventions for use in OBO Foundry ontology development. BMC Bioinformatics 2009 Apr 27;10:125 [FREE Full text] [doi: 10.1186/1471-2105-10-125] [Medline: 19397794]
8. Luo L, Xu R, Zhang GQ. Dissecting the Ambiguity of FMA Concept Names Using Taxonomy and Partonomy Structural Information. AMIA Jt Summits Transl Sci Proc 2013;2013:157-161 [FREE Full text] [Medline: 24303256]
9. Agrawal A, Perl Y, Ochs C, Elhanan G. Algorithmic detection of inconsistent modeling among SNOMED CT concepts by combining lexical and structural indicators. 2015 Nov Presented at: BIBM; 2015; Washington D.C p. 476-483. [doi: 10.1109/BIBM.2015.7359731]
10. Luo L, Mejino JLV, Zhang GQ. An analysis of FMA using structural self-bisimilarity. J Biomed Inform 2013 Jun;46(3):497-505 [FREE Full text] [doi: 10.1016/j.jbi.2013.03.005] [Medline: 23557711]

11. Luo L, Tong L, Zhou X, Mejino JLV, Ouyang C, Liu Y. Evaluating the granularity balance of hierarchical relationships within large biomedical terminologies towards quality improvement. J Biomed Inform 2017 Nov;75:129-137 [FREE Full text] [doi: 10.1016/j.jbi.2017.10.001] [Medline: 28987379]
12. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013 Sep 7. URL: https://arxiv.org/abs/1301.3781 [accessed 2020-11-19]
13. Xiao H. Bert-as-service. URL: https://github.com/hanxiao/bert-as-service [accessed 2020-11-19]
14. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. IEEE 1998;86(11):2278-2324. [doi: 10.1109/5.726791]
15. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation 1997 Nov 01;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735]
16. Vapnik V. Statistical learning theory. Hoboken: Wiley; Sep 1998:401-492.
17. Pembeci I. Using Word Embeddings for Ontology Enrichment. Int J Intell Syst Appl Eng 2016 Jul 13;4(3):49-56. [doi: 10.18201/ijisae.58806]
18. Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. J Biomed Inform 2011 Feb;44(1):163-179 [FREE Full text] [doi: 10.1016/j.jbi.2010.07.006] [Medline: 20647054]
19. Zheng F, Cui L. Exploring Deep Learning-based Approaches for Predicting Concept Names in SNOMED CT. 2018 Nov 3 Presented at: BIBM; 2018; Madrid. [doi: 10.1109/bibm.2018.8621076]
20. Zheng L, Liu H, Perl Y, Geller J. Training a Convolutional Neural Network with Terminology Summarization Data Improves SNOMED CT Enrichment. AMIA Annu Symp Proc 2019;2019:972-981 [FREE Full text] [Medline: 32308894]
21. Liu H, Geller J, Halper M, Perl Y. Using Convolutional Neural Networks to Support Insertion of New Concepts into SNOMED CT. AMIA Annu Symp Proc 2018;2018:750-759 [FREE Full text] [Medline: 30815117]
22. Liu H, Perl Y, Geller J. Transfer Learning from BERT to Support Insertion of New Concepts into SNOMED CT. AMIA Annu Symp Proc 2019;2019:1129-1138. [Medline: 32308910]
23. Foundational Model of Anatomy. URL: https://bioportal.bioontology.org/ontologies/FMA [accessed 2020-11-19]
24. Harris S, Seaborne A. SPARQL 1.1 Query Language. URL: https://www.w3.org/TR/sparql11-query/ [accessed 2020-11-19]
25. Openlink Software. URL: https://virtuoso.openlinksw.com [accessed 2020-11-19]
26. Kim Y. Convolutional Neural Networks for Sentence Classification. 2014 Oct 25 Presented at: the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014; Doha p. 1746-1751. [doi: 10.3115/v1/d14-1181]
27. Keras Documentation. URL: https://keras.io/ [accessed 2020-11-19]
28. The Stanford Parser. URL: https://nlp.stanford.edu/software/lex-parser.shtml [accessed 2020-11-19]
29. SNOMED International Homepage. URL: http://www.snomed.org [accessed 2020-11-19]
30. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer NP, et al. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. Biocomputing 2020 2019 Aug 20 [FREE Full text] [doi: 10.1142/9789811215636_0027]
31. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
32. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]

Abbreviations

Bi-LSTM: Bidirectional Long Short-Term Memory Network
CNN: Convolutional Neural Networks
FMA: the Foundational Model of Anatomy
ML: machine learning
OL: ontology learning
OWL: Web Ontology Language
PCS: parallel concept set
RDF: resource description frame
SNOMED CT: systematized nomenclature of medicine-clinical terms
SVM: support vector machine
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 10.07.20; peer-reviewed by R Abeysinghe, S Zhang; comments to author 28.07.20; revised version received 11.08.20; accepted 29.10.20; published 25.11.20.

Please cite as:
Luo L, Feng J, Yu H, Wang J
Automatic Structuring of Ontology Terms Based on Lexical Granularity and Machine Learning: Algorithm Development and Validation
JMIR Med Inform 2020;8(11):e22333
URL: http://medinform.jmir.org/2020/11/e22333/
doi:10.2196/22333
PMID:33127601

Original Paper

Parental Experiences of the Pediatric Day Surgery Pathway and the Needs for a Digital Gaming Solution: Qualitative Study

Arja Rantala[1*], RN, MHSc; Miia M Jansson[2*], RN, PhD; Otto Helve[3*], MD, PhD; Pekka Lahdenne[4*], MD, PhD; Minna Pikkarainen[5*], PhD; Tarja Pölkki[6,7,8*], RN, PhD

[1]Research Group of Medical Imaging, Physics and Technology, Research Unit of Nursing Science and Health Management, Faculty of Medicine, University of Oulu, Oulu, Finland

[2]Research Group of Medical Imaging, Physics and Technology, University of Oulu, Oulu, Finland

[3]Pediatric Research Center, Department of Pediatrics, Helsinki University Hospital, University of Helsinki, Helsinki, Finland

[4]Department of Pediatrics, Helsinki University Hospital, Helsinki, Finland

[5]Research Group of Medical Imaging, Physics and Technology, VTT Technical Research Centre of Finland, University of Oulu, Oulu, Finland

[6]Research Unit of Nursing Science and Health Management, Faculty of Medicine, University of Oulu, Oulu, Finland

[7]Department of Children and Women, Oulu University Hospital, Oulu, Finland

[8]Medical Research Center Oulu, Oulu, Finland

[*]all authors contributed equally

Corresponding Author:
Arja Rantala, RN, MHSc
Research Group of Medical Imaging, Physics and Technology
Research Unit of Nursing Science and Health Management, Faculty of Medicine
University of Oulu
Pentti Kaiteran katu 1
Oulu, FI-90014
Finland
Phone: 358 504340424
Email: arja.rantala@oulu.fi

Abstract

Background: The parents of hospitalized children are often dissatisfied with waiting times, fasting, discharge criteria, postoperative pain relief, and postoperative guidance. Parents' experiences help care providers to provide effective, family-centered care that responds to parents' needs throughout the day surgery pathway.

Objective: The objective of our study was to describe parental experiences of the pediatric day surgery pathway and the needs for a digital gaming solution in order to facilitate the digitalization of these pathways.

Methods: This was a descriptive qualitative study. The participants (N=31) were parents whose children were admitted to the hospital for the day surgical treatments or magnetic resonance imaging. The data were collected through an unstructured, open-ended questionnaire; an inductive content analysis was conducted to analyze the qualitative data. Reporting of the study findings adheres to the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist.

Results: Parental experiences of the children's day surgery pathway included 3 main categories: (1) needs for parental guidance, (2) needs for support, and (3) child involved in his or her own pathway (eg, consideration of an individual child and preparation of child for treatment). The needs for a digital gaming solution were identified as 1 main category—the digital gaming solution for children and families to support care. This main category included 3 upper categories: (1) preparing children and families for the day surgery via the solution, (2) gamification in the solution, and (3) connecting people through the solution.

Conclusions: Parents need guidance and support for their children's day surgery care pathways. A digital gaming solution may be a relevant tool to support communication and to provide information on day surgeries. Families are ready for and are open to digital gaming solutions that provide support and guidance and engage children in the day surgery pathways.

(JMIR Med Inform 2020;8(11):e23626) doi:10.2196/23626

KEYWORDS

anxiety; children; day surgery; delivery of health care; digital solution; gamification; nursing; pain; qualitative study; technology

Introduction

Background

The Global Observatory for eHealth has defined mobile health (mHealth) solutions as "medical and public health practice supported by mobile devices, such as mobile phones, patient-monitoring devices, personal digital assistants (PDAs), and other wireless devices" [1]. These digital or connected health services or solutions are changing medical and public health practices [2]. However, assessment frameworks should respect the needs and capacity of each medical system or country [3]. Gaming and gamification are areas of mHealth, which can expand users' acquiescence with health interventions and improve users' capability to self-administer and adherence to treatment [4]. Among health care professionals, developing games should be evidence-based and targeted to those who need them [5]. Gamification includes game design elements in a nongame context [6]. Game elements include competition under rules, a narrative context, a feedback and communication system, and time pressure. In addition, gamification engages users with unparalleled intensity and as per a duration system [6,7]. Serious health games (SHGs) can include educational and physiological elements as well as additional knowledge on how to influence the goal achieved through the game, such as alleviating pain or anxiety among children or parents [5,8].

For children, day surgery procedures are more common than inpatient procedures [9]. Every year, approximately 3%-10% of children (age <17 years) experience hospital stays in developed economies, such as the United States and Europe [9-11]. Day surgery patients are admitted, operated on, and discharged on the same calendar day [12,13]. A day surgical pathway should be based on a well-planned protocol, defining all the steps from planned treatment, to hospital admission, and to hospital discharge within the same day. A day surgical pathway is cost-effective and includes a nurse - delivered preanesthetic assessment consisting of preoperative education and appropriate postoperative care guidance. Such careful patient guidance reduces hospital readmissions. [14]. Surgical conditions constitute a significant proportion of the global burden of diseases [15]. However, careful patient selection, refinements in surgical and anesthetic techniques and devices, and successful patient outcomes have contributed to the optimal utilization of surgical processes along with early discharge [14].

Outpatient care also includes patients who are sedated for magnetic resonance imaging (MRI) and are discharged after treatment [16]. Guidelines from the Association of Anaesthetists and the British Association of Day Surgery underline the importance of instructions concerning day surgery and its flow [14]. An ideal day surgery pathway includes minimized waiting times, information on day surgery procedures, routine preoperative checks on the day of admission, analgesia and other postoperative information at discharge, and eventually, instructions for telephone follow-up [13]. The information provided should be of high quality, procedure-specific (for families), and age-appropriate (for children) [14,17]. Patients should be admitted to the day surgery units as close as possible to the time of their surgery.

This paper focuses on parental experiences from the pediatric day surgery pathway. Our study is part of the Icory (Intelligent Customer-driven Solution for Orthopaedic and Paediatric Surgery Care) project in which the ecosystem of hospitals, researchers, and technology providers, together with children and families, are codeveloping a digital gaming solution for the day surgery path.

According to the previous meta-analysis, digital gaming solutions can reduce children's preoperative anxiety and increase parental satisfaction [18]. In addition, digital gaming solutions can be considered as nonpharmacological distraction tools for children. Parents' experiences may help care providers to deliver more effective, family-centered care that responds to parents' needs throughout the day surgery pathway [18]. Family-centered care considers the individual needs of parents and children [19] and emphasizes the role of written information intended to reduce parental anxiety and stress regarding day surgeries [20,21]. In the child-centered care, a child is a person with their own voice and joint participation and partnership in a care environment and is considered competent [19].

In day surgery processes, appropriate preoperative preparation is crucial [22,23]. Usually, the day surgical treatment is a unique experience for both the child and family. The incidence of preoperative anxiety in children varies between 40% and 75% [24,25]. In order to avoid unexpected stress, the whole family needs to be prepared for upcoming surgery [26]. Untreated anxiety is associated with increased intensity of pain afterwards [26-28]. In addition, parental anxiety has an enormous effect on children's preoperative anxiety, which correlates with increased postoperative pain in children [28]. Correspondingly, the factor of postoperative pain is associated with parental satisfaction in pediatric day surgery [29].

The parents of hospitalized children perceive high levels of stress and anxiety [30,31]. In addition, they are dissatisfied with waiting times [32,33], fasting [32], discharge criteria [13,34], and postoperative pain relief [29]. Parental satisfaction with treatment and the care itself, however, is good [29].

Our study focuses on the gap between the information provided to parents and their needs. In addition, the expectations from a digital gaming solution were addressed to support the digitalization of pediatric day surgery pathways. The research questions were as follows:

1. What are the experiences of parents of care in the pediatric day surgery pathway?
2. What are the needs of parents for a digital gaming solution in the pediatric day surgery pathway?

Methods

Design

This was a descriptive qualitative study based on the experiences of parents of hospitalized children [35,36].

Participants and Settings

Participants were asked to volunteer and participate in the study using convenience sampling [35]. The inclusion criteria were as follows: parent or custodian of a child who was receiving a

day surgical treatment at the selected hospital, ability to understand and write in Finnish, and access to a laptop or a mobile app for answering the questionnaire. The selected participants were acquainted with the research topic [37] because of the experience of their child's day surgery pathway in the hospital.

The respondents consisted of parents (N=31) whose children were admitted to otolaryngologic surgery (n=7), plastic surgery (n=3), oral/dental surgery (n=1), ophthalmic surgery (n=1), orthopedic surgery (n=7), soft tissue surgery (n=3), gastroenterological surgery (n=5), vascular surgery (n=1), and MRI (n=3). Out of the 31 parents, 23 (74%) were females aged 30-39 years, and 8 (26%) were males.

The study was conducted at the university hospital in Finland, a 140-bed tertiary care pediatric hospital. The hospital provides specialized health care in pediatrics, including pediatric surgery, child neurology, and child psychiatry. In addition, the hospital has been assigned the national treatment responsibilities in specific pediatric conditions, for example, cardiac surgery and organ transplant. In 2019, a total of 6883 surgical operations were performed of which 3300 were day surgery procedures.

Data Collection

The data were collected using an unstructured questionnaire. The questionnaire was planned and designed via remote meetings in 2018 by a panel of 7 specialists from the University of Oulu, VTT Technical Research Centre of Finland, Helsinki University Hospital, and Oulu University Hospital. The panel included pediatricians, nurses, and researchers. The questionnaire was designed to gather information on current pathways of children's day surgeries in order to develop a more digitized pathway; it was based on the findings from an earlier study [2]. The questionnaire was designed to address the research questions using 3 demographic and 6 open-ended questions (Textbox 1). The questionnaire was tested in March 2019 by a PhD student and a professor in the selected hospital with parents who were waiting for treatment for their children. The parents were invited to respond to Questback Essentials' web-based unstructured questionnaire via a quick response (QR) code using their own mobile phones. A total of 5 parents responded, and thus the questionnaire was considered usable.

Textbox 1. Unstructured questionnaire for parents.

1. Respondents' demographics
2. Respondents' demographics
3. Has your child been in a day surgery unit? Yes/No
4. What kind of treatment has your child had?
5. What kind of information did you receive from the hospital staff about what is going to happen during your child's day surgery treatment?
6. Could you describe what kind of support you received during your child's treatment?
7. Can you tell us how your child was involved in the care path provided by the hospital staff?
8. Did you have an access to any treatment-related games during your child's care that you would have played with your children?
9. Demo of the Icory solution: What do you think of this solution? What kind of solution should be available?

Data collection was conducted from October 2019 to December 2019. The research nurse recruited voluntary respondents from the hospital's recovery room while children were in the postanesthesia care unit. The parents answered the anonymous questionnaire via their own smartphones or laptops. If the respondent did not have a smartphone or a laptop, one was provided to them by the research nurse. The researchers were not able to identify individual respondents.

Data Analysis

The qualitative data were analyzed by conducting an inductive content analysis, which included 3 main phases: preparation, organization, and reporting [35,36]. Inductive content analysis is used when knowledge is fragmented or when there is little knowledge of the phenomenon [38]. Content analysis is a method of analyzing written, verbal, or visual communication that distills words into content-related categories for providing new knowledge of the phenomenon [38,39]. The data were analyzed using NVivo 12 (QRS International), a qualitative research software. Demographic data were reported using frequencies and percentages.

First, the data were evaluated for quality by 3 researchers (AR, MMJ, and TP). The data were transferred to NVivo 12. In the analysis, initial impressions were written down as notes. Second, the data were abstracted into open codes and transferred into tables. Similarly, open codes were grouped into categories [38]. Third, those categories were named using a word that was characteristic of the content and were formed into subcategories. Similar subcategories were grouped together and called upper categories. The process produced 294 open codes, 21 subcategories, 9 upper categories, and 4 main categories. The data analysis was conducted by 2 researchers (AR and MMJ), discussed and agreed upon by 3 researchers (AR, MMJ, and TP), and commented on by the researchers who developed the questionnaire (OH, MP, and PL).

Rigor

Rigor was ensured using the criteria of credibility, dependability, confirmability, transferability, and authenticity [35,40,41]. Credibility and dependability in this study were ensured by designing the questionnaire based on early studies [2,25,42] and based on the actual needs of the selected hospital. Confirmability was established in the analysis process through

researcher triangulation (AR, MMJ, and TP). The preparation phase included defining participant inclusion criteria, planning the data collection in questionnaire format to reach data saturation, and selecting the unit of analysis.

Saturation was achieved when the written answers began to repeat themselves in the collected data. Transferability was established in this study by describing the hospital environment in which the study was conducted in order to enable the interpretation and transfer of the results in other contexts. Authenticity was ensured through quotations to indicate the richness of data [41]. The reporting was performed systematically according to the COREQ (Consolidated Criteria for Reporting Qualitative Research) criteria (Multimedia Appendix 1) [43].

Ethical Considerations

The study was reviewed by the local ethical committee (decision #3181-2018) and was granted a research permit (decision #284-2019). The aim of this study was verbally explained to the respondents by the research nurse, and they were informed that responding to the questionnaire was entirely voluntary. The questionnaire was designed so that it would be impossible to identify the respondents. The study followed the Helsinki Declaration [44]. All participants were informed about the voluntary nature of the research [21,45].

Results

Analysis of the data for the first research question revealed 3 main categories related to the parents' experiences of their children's day surgery pathways: (1) needs for parental guidance (which included 2 upper categories—content of information and patient flow during the day surgery pathway), (2) needs for support in the children's day surgery pathway (which included the upper categories—physiological support for children and psychological support for children and families), and (3) child involvement in their own pathway (which included upper categories—consideration of an individual child and preparing a child for treatment) (Table 1).

Table 1. Parental experiences of the care in the children's day surgery pathway.

Categories	Experiences
Main category 1: Needs for parental guidance	
Upper category 1: Content of information	
Subcategory 1: Transparency of the pathway	• Lack of knowledge regarding operating theater activities • Unexpected follow-up overnight • Unexpected changes in day surgery pathway • Lack of knowledge regarding day surgery pathway • Lack of instruction regarding surgical wound • Instructions could be sent by email • Instructions in the information letter should be updated
Subcategory 2: Analgesia and amnesia	• Lack of knowledge regarding pain management • Need for pain management guidance • Need for information about anesthesia
Upper category 2: Patient flow during the day surgery pathway	
Subcategory 1: Physical environment	• Lack of information regarding free parking • Guidance signs are needed • Timely permission for entering the waiting room
Subcategory 2: Waiting time	• The overall waiting time was too long • Waiting is challenging with hungry children • The waiting time for the operation was too long • Progress regarding the waiting time should be shared by nurse • Unsustainable schedules
Subcategory 3: Roles and responsibilities	• Doctor did not call following operation • Discrepancies in responsibilities • Superficiality of information • Nurses are better at sharing information
Main category 2: Needs for support	
Upper category 1: Physiological support for children	
Subcategory 1: Eating after operation	• Meal requirements • Unsuitable food after operation • Conflicting information regarding available meals
Subcategory 2: Environment safety	• Hearing protectors for sound-sensitivity
Upper category 2: Psychological support for children and family	
Subcategory 1: Rewards and trophies	• Rewards are pleasing
Subcategory 2: Psychological needs	• Parents' psychological needs should be enquired about • Parental well-being should be considered
Subcategory 3: Timing of guidance	• Correct timing for nursing guidance • Parental role and timing in sudden situation
Main category 3: Child's involvement in his or her own pathway	
Upper category 1: Consideration of individual children	
Subcategory 1: Individual needs	• Needs of sound-sensitive children • Own devices for sound-sensitive child
Subcategory 2: Giving more time to children	• More time should be given to children by the nurse • Calming should be ensured before operation
Upper category 2: Preparation for treatment	
Subcategory 1: Preparation for treatment	• Familiarization with treatment should be ensured

Categories	Experiences
Subcategory 2: Preparation with games and pictures	• Pictures to help preparation • Games to help preparation

Analysis of the data for the second research question revealed 1 main category, a digital gaming solution for children and families to support care. This category included 3 upper categories: preparing children and families for the day surgery via the solution, gamification in the solution, and connecting people through the solution (Table 2).

Table 2. Parental needs for a gamification solution in the day surgery pathway.

Digital gaming solution for children and families to support care	Experiences
Upper category 1: Preparing children and families for the day surgery via the solution	
Subcategory 1: Preparing via digital gaming solution	• General information about surgery • Instructions available beforehand in solution
Subcategory 2: Virtual familiarization with the care environment	• Virtual tour • Familiarization with operation room beforehand • Visibility of real environment
Subcategory 3: Waiting time in solution	• Information on waiting time • Distraction from waiting
Upper category 2: Gamification in the solution	
Subcategory 1: Gamification to overcome hospital anxiety and fear	• Games with music • Videos in solution • Fun in order to relieve fear • Games for reducing fear
Subcategory 2: Gamification in support of care	• Games for preparation • Games for rehabilitation • Age-appropriate material for children
Upper category 3: Connecting people through the solution	
Subcategory 1: Interaction between the medical staff, families, and children	• Interaction between nurses and patient • Support and guidance from staff (apart from game
Subcategory 2: Peer support	• Children telling stories about treatment via solution • Ability to share feelings

Needs for Parental Guidance

The parents described their experiences and ideas regarding the information that could help both families and children address the challenges they experienced in the day surgery pathway. Identified categories were related to the content of information and patient flow during the day surgery pathway.

Content of Information

This upper category included 2 subcategories: transparency of the pathway and analgesia and amnesia. Generally, the content of the provided information (eg, admission instructions, dining, fasting, overnight stay, pain management, parking, patient flow—including preoperative preparation, time, and place) was considered sufficient. For instance, one parent wrote:

> *I think we got proper information throughout the whole treatment process. They told us where to go and how to prepare for the treatment.*

However, some respondents faced challenges regarding the transparency of the pathway and analgesia and amnesia, which are subcategories of this upper category.

Transparency of the pathway included the needs for knowledge of information concerning the children's day surgical pathway. Parents faced challenges in receiving information on the stages, milestones, and procedures of day surgeries. According to parents, there was a lack of knowledge about the discharge criteria. Moreover, sudden changes brought further challenges in obtaining information. For instance, one respondent wrote: "It was supposed to be a day surgery treatment, but our child needed to be monitored overnight." Correspondingly, respondents made suggestions related to the information on postsurgery care: "I would like to have better instructions on what to be aware of and what not to do with a surgical wound."

According to parents, a digital solution with different kinds of features (eg, email) could be utilized in order to enhance information transfer, as one of the respondents remarked, "Yeah, an email could have been a working solution" for providing

information about the day surgery. The implementation of information transfer (eg, individual counseling delivered via face-to-face contact and telephone; written counseling delivered via letter, email, and SMS reminders) was considered sufficient. One of the parents expressed the following view: "The flow of the care pathway was well communicated by the medical staff on the day, before the treatment, and before and after the treatment in the recovery room."

Analgesia and amnesia included needs for information concerning pain management. According to respondents, there was a lack of information on the management of postoperative pain and treatment-related pain. One of the respondents maintained, "We don't know anything about that (pain management at home) yet, I have asked about possible pain but..." In addition, there is a lack of knowledge related to amnesia/sedation. For instance, one respondent wrote:

> *The one thing we were worried about was how amnesia would go and we asked about it. However, that was confirmed in the operation room.*

Patient Flow During the Day Surgery Pathway

The parents described their experiences and ideas regarding the patient flow in the day surgery pathway. Identified categories were related to the physical environment, waiting time, and roles and responsibilities, which are 3 subcategories of this upper category.

Physical environment was related to the lack of information regarding free parking and guidance signs. The respondents faced challenges in accessing parking and parents' waiting areas. According to the respondents, more information about the free car parking and waiting places needed to be added to the information letter. One of the respondents described the following challenge:

> *We haven't been to the place (hospital) before. I verified the location of the car park from the website.*

The respondent also added: "The only thing was that the request for entry to the parents' living room was made too late."

Waiting time was related to the main challenges and needs. The arrival time, for instance, was considered to be the same as surgery time. One of the respondents expressed this view by stating, "We had time for the treatment but still we had to wait with other parents for 2-3 hours." Overall, the waiting time was considered too long, and more information regarding the remaining waiting time was warranted. The following excerpt from one of the respondents expresses such a view: "There should have been some kind of information about how long our waiting time (for the treatment) would be." The timing of patient counseling was also considered nonoptimal in certain circumstances.

Roles and responsibilities were related to the challenges faced by children and parents in their interactions with the hospital staff. It was unclear how and who would announce what. Despite certain promises, surgeons did not share information about the surgery before and after surgery, or they shared information about surgery superficially and briefly. For instance, one respondent wrote, "There was information on the screen in the waiting room that the doctor would call after the treatment, but this never happened."

In addition, the answers received were somewhat indicative/suggestive. However, the respondents made some suggestions regarding the digital gaming solutions and child involvement in their own care pathways.

Needs for Support

Support for Children and Family

The second main category included 2 upper categories that parents described as needs for the day surgical pathway: physiological support for children and psychological support for children and families. The received support (eg, explanations, parents' involvement, support from nearby, and friendliness of staff) was considered sufficient. In addition, most of the respondents felt that the service provided by the hospital was friendly, attentive, and informative.

Physiological Support

This upper category was related to the challenges of eating after an operation and environmental safety. Conflicting information was observed related to postsurgery meals. The postoperation meal requirements were referenced in many responses, as was the need for hearing protectors for sound-sensitive children.

Psychological Support for Children and Families

This upper category addressed the need for rewarding children after operation, timing of guidance from nurses in sudden situations, parental role and timing in sudden situations, and lack of psychological support for parents. Parental involvement was also related to the challenges in their roles. One respondent stated, "It was hard in the operating room (before anesthesia), when you should be focused on your child's excitement and at the same time matters relating to anaesthesia."

One of the respondents felt that their psychological needs were ignored in the hospital and mentioned, "Psychological needs were not taken into account and they (hospital staff) could have asked us about it."

Child Involvement in His or Her Own Care Pathway

Consideration and Preparation

The parents described their experiences and ideas regarding their children's involvement in their own care pathways. The identified categories were related to the consideration of an individual child and preparation for treatment, reflecting 2 upper categories in this main category.

Children's involvement in their own care pathway (eg, answering the child's questions, talking to the child, turning attention elsewhere, considering the child's fear, encouraging and praising the child, giving time to situations faced by children, listening to the child, involving the child in the care path, and taking the individual into account) was considered sufficient. For instance, one respondent stated, "My child was also allowed to ask questions that bothered him, and they were answered really well."

However, some respondents had faced challenges regarding consideration of an individual child and preparing a child for treatment, which were subcategories for this main category.

Consideration of an Individual Child

This category was related to the needs of children with special needs and allowing time for children to calm down. The individual needs of sound-sensitive children should be taken into account.

Preparation for Treatment

This category included the challenges related to children's needs for parents in the operation room before anesthesia, needs to familiarize themselves with the hospital environment in advance, and needs of requiring more time to deal with frightening situations. Respondents collectively described these needs as giving time to children before their operations. One respondent explained this need as follows:

> *Going to the operating room caused a little extra stress, as the speed was faster there than the speed in the restroom. It would have been good to calm the child down a bit before starting the operation.*

A Digital Gaming Solution for Children and Families to Support Care

Preparation, Gamification and Connection

Respondents described their needs for a digital gaming solution that could help children and families in children's day surgery care. This included 3 upper categories: preparing children and families for the day surgery, gamification in solutions, and connecting people through the solution (Table 2).

Games are not implemented in the hospital's current pathways of pediatric day surgeries. However, parents reported positive attitudes toward a digital gaming solution for pediatric day surgery. For instance, one respondent noted, "Today's kids are born at this time of technology, so I think games could work well for kids of a certain age."

In addition, attitudes toward rehabilitation through playing and gaming were positive. The identified requirements for digital gaming solutions were divided into 7 subcategories.

Preparing Children and Families for the Day Surgery via the Solution

The proposed needs for a digital gaming solution were related to 3 different items: preparing via a digital gaming solution, virtual familiarization with the care environment, and managing the waiting time with a solution. These are the subcategories of this upper category.

Preparing via a Digital Gaming Solution

This subcategory included aspects that could enable children to prepare for their treatment. The solution should include general information about surgery, whereas instructions concerning treatment were supposed to be already available. Information storage in the game would decrease the need for information retrieval. According to respondents, the developed digital gaming solution should be easy to use. An informative gaming solution would reduce the need for Googling, as a respondent added, "That kind of solution would reduce need for Google."

Virtual Familiarization With the Care Environment

Virtual familiarization with the care environment included parental needs for a virtual tour for children and families, information about the operating room via virtual visits, and the ability to see the hospital via the solution. In addition, a digital gaming solution could include genuine pictures of various hospital spaces.

Managing the Waiting Time With a Solution

This subcategory, managing the waiting time with a solution, was seen as an important requirement for the solution. The solution would need to provide parents with information about the waiting times following the treatments. In addition, the solution could be applied in other circumstances to relieve waiting.

Gamification in the Solution

The upper category of gamification in the solution included the following 2 subcategories: gamification to overcome hospital anxiety and fear and gamification in support of care. This upper category also considers parents' expectations from gamification in the solution.

Gamification to Overcome Hospital Anxiety and Fear

This category included parental needs for solutions for children. According to parents, a digital gaming solution could include all sorts of fun and interactive features (eg, videos, games, and music) to reduce fear. The following excerpt expresses this view:

> *After all, children can't help but like everything interactive and cool. Even if the device offers nothing but fun for the child, it will certainly be helpful to relieve fear.*

Gamification in Support of Care

According to respondents, the gamification could include age-appropriate information regarding the most common types of surgeries. According to parents, certain games could be utilized for the preparation for surgery as well as postoperative recovery (eg, rehabilitation). For instance, one respondent stated, "It would be good to prepare themselves for treatment via a solution."

Connecting People Through a Solution

This upper category included 2 subcategories—interactions between medical staff, families, and children and peer support. The parents stated that the digital gaming solution could enable interaction between the hospital and families and enable the peer support: "I like the thought that with the help of the solution parents could connect with nurses."

According to parents, a digital gaming solution could include children's own positive stories regarding their day surgery pathways. In addition, a digital gaming solution could enable the sharing of feelings with others in order to reduce fear. One of the respondents expressed, "Being able to share their feelings

with others going through similar measures could be a good way to address their fears."

Discussion

Principal Results

To the best of our knowledge, this is the first qualitative study that explores parental experiences throughout the entire pathway of pediatric day surgeries in order to support the digitalization of that pathway in the selected hospital. Our findings revealed that although the current content and information transfer were considered sufficient, parents expected (1) better guidance related to the content of information, (2) more psychological support, and (3) involvement of children in their own care pathway. Additionally, it was found that there was a need for a digital gaming solution that would provide the required information and help families to be better prepared for their oncoming treatments.

Strengths and Limitations

The results present the experiences of the needs of parents in their children's day surgery pathways. Almost all (28/31, 90%) children had been in a day surgery, whereas 10% (3/31) respondents had undergone an MRI. The number of respondents was reasonably small—only 31—and the text material they produced (that was analyzed in this study) was brief, as it usually is when responding via the internet. However, the respondents had a fresh perspective on the care the hospital provided to their children, which strengthens our results.

The data could have had a greater breadth if there had been an opportunity to conduct individual or focus group interviews. Then the interviewees could have provided richer material to be analyzed, or different perspectives could have been clarified. However, the saturation was achieved, and the respondents produced texts that included rich material for our inductive content analysis. For future studies, the perspectives of children should be included for broadening and strengthening the results.

It was not possible to get feedback on the results at an organized event at the hospital because the respondents were anonymized, and they did not ask the research nurse any further questions while responding to the anonymous questionnaire. The results are transferable to similar contexts where a hospital has developed its own digital environment, but the generalizability of the results would require further quantitative research with a larger sample size.

Comparison With Prior Work

The need for support is in line with previous studies [21,23,29], which explained a situation in which parents were unfamiliar with how to conduct postoperative pain management for their children. However, in our study, the parents were ready to view a digital gaming solution as a relevant tool for supporting care in different situations concerning the children's day surgery pathways. In recent studies, parents also needed more guidance regarding fasting [23,32], equipment used in the operating and recovery rooms [21], discharge criteria [13,34], and postoperative complications [23,29]. In our study, parents wanted to increase gamification, as it was considered as an important aspect of the required guidance. The need for psychological support is in line with previous literature, in that participants were most frequently dissatisfied with waiting times [21,32,33,46]. Thus, interventions aimed at reducing waiting times and raising patient satisfaction are warranted. In our study, parents were ready for a digital gaming solution, which could be used in different kinds of situations, such as waiting for treatment, reducing children's anxiety, and patient guidance. For future studies or for developing SHGs for parents and children in the day surgery pathway, our study has addressed the first stage of developing a gaming solution. We have identified a target audience and expected outcomes [5].

According to respondents, the digital gaming solution could be used to help families be better prepared for the coming treatments. This could include a virtual tour of a hospital to familiarize children with the environment beforehand. In the study by Carlsson and Henningsson [47], the researchers realized that visiting the operation room might not reduce parents' or children's anxiety in surgery care situations. This finding is contradicted by the study of Rantala et al [18], who argued based on a recent meta-analysis that web-based interventions (eg, educational web-based programs or age-appropriate streamed videos) could be used to reduce children's anxieties [18]. In another study by Rantala et al [23], health specialists observed that a digital gaming solution (developed for a hospital environment, including virtual visits to the hospital for different surgeries) would help families and children to be better oriented to an upcoming treatment. In our study, parents considered a child's personal involvement in his or her care to be very important. A gaming digital solution developed throughout the care path can solve this challenge. In addition, the World Health Organization [1] raises an important aspect that mHealth could be used to increase patient commitment to their own care and to develop a more personalized path for patient care. In a previous systematic review, gamification was mostly used in chronic diseases and for improving physical activity among patients; there were no designed SHGs for children's day surgery pathway [4]. Thus, our study produces new knowledge for the rapid day surgery pathway and helps developers consider the parental perspective. In addition, peer support has been used in mHealth app solutions for children with chronic diseases [4,48]. In our study, parents were open to peer support via a gaming digital solution. For future studies, it would be important to examine whether this could be used as an alternative method of patient counseling that could reduce hospital readmissions in children's day surgeries [14,34].

Overall, the results of the study agree with the existing literature on patients' expectations and needs related to hip and knee arthroplasty [49]. In this study, the patients suggested that they needed a digital solution that allows for better real-time communication methods (eg, information transfer, discussion forums) and patient counseling (eg, resources, content and implementation). In addition, our results support the previous literature on treatment pathways for surgical patients [49-52]. To the best of our knowledge, prior studies have not focused on the pathways of pediatric day surgeries. In this respect, our study produces new research.

Conclusion

The parents of children in day surgeries need reliable information about the pathways of those surgeries. Children must be involved in the care paths of their day surgeries. These parents were open to digital gaming solutions. They expressed their thoughts on what kinds of solutions would be relevant to the clinical practice and could positively affect the care provision in hospitals. The digital gaming solution should be developed for the needs of children and provide important information about day surgeries to families. The findings of this study can be applied for integrating digital solutions into hospital environments.

Acknowledgments

The authors wish to thank all the parents and participants who shared their experiences by responding to our questionnaire. The research was conducted as part of the project entitled Icory (Intelligent Customer-driven Solution for Orthopaedic and Paediatric Surgery Care), which was funded by Business Finland, a Finnish Funding agency, during 2018-2020. The funder has not influenced the design, conduct, analysis, or reporting of the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist for qualitative research.
[PDF File (Adobe PDF File), 193 KB - medinform_v8i11e23626_app1.pdf]

References

1. World Health Organization. mHealth: New Horizons for Health Through Mobile Technologies: Second Global Survey on eHealth. Geneva: World Health Organization; 2011. URL: https://apps.who.int/iris/bitstream/handle/10665/44607/9789241564250_eng.pdf [accessed 2020-10-22]
2. Niemelä R, Pikkarainen M, Ervasti M, Reponen J. The change of pediatric surgery practice due to the emergence of connected health technologies. Technological Forecasting and Social Change 2019 Sep;146:352-365 [FREE Full text] [doi: 10.1016/j.techfore.2019.06.001]
3. Bradway M, Carrion C, Vallespin B, Saadatfard O, Puigdomènech E, Espallargues M, et al. mHealth Assessment: Conceptualization of a Global Framework. JMIR Mhealth Uhealth 2017 May 02;5(5):e60 [FREE Full text] [doi: 10.2196/mhealth.7291] [Medline: 28465282]
4. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform 2017 Jul;71:31-48 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.011] [Medline: 28536062]
5. Verschueren S, Buffel C, Vander Stichele G. Developing Theory-Driven, Evidence-Based Serious Games for Health: Framework Based on Research Community Insights. JMIR Serious Games 2019 May 02;7(2):e11565 [FREE Full text] [doi: 10.2196/11565] [Medline: 31045496]
6. Zichermann G, Cunningham C. Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps. Sebastopol: O'Reilly Media Inc; 2011.
7. Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness. New York: Association for Computing Machinery; 2011 Sep Presented at: Proceedings of the 15th International Academic MindTrek Conference; Sep 28, 2011; Tampere, Finland p. 9-15 URL: https://doi.org/10.1145/2181037.2181040 [doi: 10.1145/2181037.2181040]
8. Buffel C, van Aalst J, Bangels A, Toelen J, Allegaert K, Verschueren S, et al. A Web-Based Serious Game for Health to Reduce Perioperative Anxiety and Pain in Children (CliniPup): Pilot Randomized Controlled Trial. JMIR Serious Games 2019 Jun 01;7(2):e12431 [FREE Full text] [doi: 10.2196/12431] [Medline: 31199324]
9. Omling E, Jarnheimer A, Rose J, Björk J, Meara JG, Hagander L. Population-based incidence rate of inpatient and outpatient surgical procedures in a high-income country. Br J Surg 2018 Jan;105(1):86-95 [FREE Full text] [doi: 10.1002/bjs.10643] [Medline: 29131303]
10. Witt W, Weiss A, Elixhauser A. Overview of Hospital Stays for Children in the United States, 2012. Statistical Brief #187. 2014 Dec. URL: https://hcup-us.ahrq.gov/reports/statbriefs/sb187-Hospital-Stays-Children-2012.pdf [accessed 2020-05-30]
11. Hospital discharges and length of stay statistics. Eurostat Statistics Explained. 2017. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Hospital_discharges_and_length_of_stay_statistics [accessed 2020-08-11]
12. Castoro C, Bertinato L, Baccaglini U, Drace C, McKee M, IAAS Executive Committee Members. Policy brief- Day Surgery: Making It Happen. 2007. URL: https://www.euro.who.int/__data/assets/pdf_file/0011/108965/E90295.pdf [accessed 2020-08-01]
13. Quemby D, Stocker M. Day surgery development and practice: key factors for a successful pathway. 2014 Dec;14(6):256-261 [FREE Full text] [doi: 10.1093/bjaceaccp/mkt066]

14. Bailey CR, Ahuja M, Bartholomew K, Bew S, Forbes L, Lipp A, et al. Guidelines for day-case surgery 2019: Guidelines from the Association of Anaesthetists and the British Association of Day Surgery. Anaesthesia 2019 Jun;74(6):778-792 [FREE Full text] [doi: 10.1111/anae.14639] [Medline: 30963557]
15. Shrime MG, Bickler SW, Alkire BC, Mock C. Global burden of surgical disease: an estimation from the provider perspective. Lancet Glob Health 2015 Apr 27;3 Suppl 2:S8-S9 [FREE Full text] [doi: 10.1016/S2214-109X(14)70384-5] [Medline: 25926322]
16. Deen J, Vandevivere Y, Van de Putte P. Challenges in the anesthetic management of ambulatory patients in the MRI suites. Curr Opin Anaesthesiol 2017 Dec;30(6):670-675. [doi: 10.1097/ACO.0000000000000513] [Medline: 28817401]
17. EACH European Association for Children in Hospital. The EACH Charter with Annotations. 2016. URL: https://www.each-for-sick-children.org/images/stories/2016/Charter_AUG2016_oSz.pdf [accessed 2020-10-29]
18. Rantala A, Pikkarainen M, Miettunen J, He H, Pölkki T. The effectiveness of web-based mobile health interventions in paediatric outpatient surgery: A systematic review and meta-analysis of randomized controlled trials. J Adv Nurs 2020 Apr 13. [doi: 10.1111/jan.14381] [Medline: 32281673]
19. Coyne I, Holmström I, Söderbäck M. Centeredness in Healthcare: A Concept Synthesis of Family-centered Care, Person-centered Care and Child-centered Care. J Pediatr Nurs 2018;42:45-56. [doi: 10.1016/j.pedn.2018.07.001] [Medline: 30219299]
20. Landier M, Villemagne T, Le Touze A, Braïk K, Meignan P, Cook AR, et al. The position of a written document in preoperative information for pediatric surgery: A randomized controlled trial on parental anxiety, knowledge, and satisfaction. J Pediatr Surg 2018 Mar;53(3):375-380. [doi: 10.1016/j.jpedsurg.2017.04.009] [Medline: 28456425]
21. Healy K. A descriptive survey of the information needs of parents of children admitted for same day surgery. J Pediatr Nurs 2013 Apr;28(2):179-185. [doi: 10.1016/j.pedn.2012.07.010] [Medline: 22892072]
22. Yahya Al-Sagarat A, Al-Oran HM, Obeidat H, Hamlan AM, Moxham L. Preparing the Family and Children for Surgery. Crit Care Nurs Q 2017;40(2):99-107. [doi: 10.1097/CNQ.0000000000000146] [Medline: 28240692]
23. Rantala A, Pikkarainen M, Pölkki T. Health specialists' views on the needs for developing a digital gaming solution for paediatric day surgery: A qualitative study. J Clin Nurs 2020 Sep;29(17-18):3541-3552. [doi: 10.1111/jocn.15393] [Medline: 32614105]
24. Chorney JM, Kain ZN. Behavioral analysis of children's response to induction of anesthesia. Anesth Analg 2009 Nov;109(5):1434-1440. [doi: 10.1213/ane.0b013e3181b412cf] [Medline: 19713262]
25. Cumino DO, Vieira JE, Lima LC, Stievano LP, Silva RAP, Mathias LAST. Smartphone-based behavioural intervention alleviates children's anxiety during anaesthesia induction: A randomised controlled trial. Eur J Anaesthesiol 2017 Mar;34(3):169-175. [doi: 10.1097/EJA.0000000000000589] [Medline: 28146459]
26. Rabbitts JA, Groenewald CB, Tai GG, Palermo TM. Presurgical psychosocial predictors of acute postsurgical pain and quality of life in children undergoing major surgery. J Pain 2015 Mar;16(3):226-234 [FREE Full text] [doi: 10.1016/j.jpain.2014.11.015] [Medline: 25540939]
27. Kain ZN, Mayes LC, Caldwell-Andrews AA, Karas DE, McClain BC. Preoperative anxiety, postoperative pain, and behavioral recovery in young children undergoing surgery. Pediatrics 2006 Aug;118(2):651-658. [doi: 10.1542/peds.2005-2920] [Medline: 16882820]
28. Fortier MA, Del Rosario AM, Martin SR, Kain ZN. Perioperative anxiety in children. Paediatr Anaesth 2010 Apr;20(4):318-322. [doi: 10.1111/j.1460-9592.2010.03263.x] [Medline: 20199609]
29. Sam CJ, Arunachalam PA, Manivasagan S, Surya T. Parental Satisfaction with Pediatric Day-Care Surgery and its Determinants in a Tertiary Care Hospital. J Indian Assoc Pediatr Surg 2017;22(4):226-231 [FREE Full text] [doi: 10.4103/jiaps.JIAPS_212_16] [Medline: 28974875]
30. Commodari E. Children staying in hospital: a research on psychological stress of caregivers. Ital J Pediatr 2010 May 25;36:40 [FREE Full text] [doi: 10.1186/1824-7288-36-40] [Medline: 20500854]
31. Scrimin S, Haynes M, Altoè G, Bornstein MH, Axia G. Anxiety and stress in mothers and fathers in the 24 h after their child's surgery. Child Care Health Dev 2009 Mar;35(2):227-233 [FREE Full text] [doi: 10.1111/j.1365-2214.2008.00920.x] [Medline: 19228156]
32. Elebute O, Ademuyiwa A, Seyi-olajide J, Bode C. An audit of parental satisfaction of pediatric day case surgery at the Lagos University Teaching Hospital. J Clin Sci 2014;11(2):44. [doi: 10.4103/1595-9587.146501]
33. Criss CN, Brown J, Gish JS, Gadepalli SK, Hirschl RB. Clinic-day surgery for children: a patient and staff perspective. Pediatr Surg Int 2018 Jul;34(7):755-761. [doi: 10.1007/s00383-018-4288-3] [Medline: 29808282]
34. Short HL, Parakati I, Heiss KF, Wulkan ML, Sweeney JF, Raval MV. Challenge of balancing duration of stay and readmissions in children's operation. Surgery 2017 Oct;162(4):950-957. [doi: 10.1016/j.surg.2017.06.005] [Medline: 28709646]
35. Polit D, Beck C. Nursing Research: Generating and Assessing Evidence For Nursing Practice. 10th ed. Philadelphia: Wolters Kluwer; 2017.
36. Elo S, Kääriäinen M, Kanste O, Pölkki T, Utriainen K, Kyngäs H. Qualitative Content Analysis. SAGE Open 2014 Feb 11;4(1):215824401452263. [doi: 10.1177/2158244014522633]

37. Kyngäs H, Elo S, Pölkki T, Kääriäinen M, Kanste O. Sisällönanalyysi suomalaisessa hoitotieteellisessä tutkimuksessa (The use of content analysis in Finnish nursing science research). Hoitotiede 2011;23(2):138-148 [FREE Full text]
38. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008 Apr;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
39. Cole FL. Content analysis: process and application. Clin Nurse Spec 1988;2(1):53-57. [doi: 10.1097/00002800-198800210-00025] [Medline: 3349413]
40. Lincoln YS, Guba EG, Pilotta JJ. Naturalistic inquiry. International Journal of Intercultural Relations 1985 Jan;9(4):438-439. [doi: 10.1016/0147-1767(85)90062-8]
41. Guba EG, Lincoln Y. Competing paradigms in qualitative research. In: Denzin NK, Guba EG, editors. Handbook of Qualitative Research. Thousand Oaks, CA: Sage; 1994:105-117.
42. Fortier MA, Bunzli E, Walthall J, Olshansky E, Saadat H, Santistevan R, et al. Web-based tailored intervention for preparation of parents and children for outpatient surgery (WebTIPS): formative evaluation and randomized controlled trial. Anesth Analg 2015 Apr;120(4):915-922 [FREE Full text] [doi: 10.1213/ANE.0000000000000632] [Medline: 25790213]
43. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
44. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
45. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
46. Hui WJ, Pikkarainen M, Nah SA, Nah SNJ, Pölkki T, Wang W, et al. Parental Experiences While Waiting For Children Undergoing Surgery in Singapore. J Pediatr Nurs 2020;52:e42-e50. [doi: 10.1016/j.pedn.2020.01.004] [Medline: 31983480]
47. Carlsson RNE, Henningsson RN. Visiting the Operating Theatre Before Surgery Did Not Reduce the Anxiety in Children and Their Attendant Parent. J Pediatr Nurs 2018;38:e24-e29. [doi: 10.1016/j.pedn.2017.09.005] [Medline: 28939000]
48. Cafazzo J, Casselman M, Hamming N, Katzman DK, Palmert MR. Design of an mHealth app for the self-management of adolescent type 1 diabetes: a pilot study. J Med Internet Res 2012 May 08;14(3):e70 [FREE Full text] [doi: 10.2196/jmir.2058] [Medline: 22564332]
49. Jansson MM, Harjumaa M, Puhto A, Pikkarainen M. Healthcare professionals' proposed eHealth needs in elective primary fast-track hip and knee arthroplasty journey: A qualitative interview study. J Clin Nurs 2019 Dec;28(23-24):4434-4446. [doi: 10.1111/jocn.15028] [Medline: 31408555]
50. Jansson MM, Harjumaa M, Puhto A, Pikkarainen M. Patients' satisfaction and experiences during elective primary fast-track total hip and knee arthroplasty journey: A qualitative study. J Clin Nurs 2020 Feb;29(3-4):567-582. [doi: 10.1111/jocn.15121] [Medline: 31769559]
51. Jansson M, Koivisto J, Pikkarainen M. Identified opportunities for gamification in the elective primary fast-track total hip and knee arthroplasty journey: Secondary analysis of healthcare professionals' interviews. J Clin Nurs 2020 Jul;29(13-14):2338-2351. [doi: 10.1111/jocn.15246] [Medline: 32222001]
52. Jansson MM, Harjumaa M, Puhto A, Pikkarainen M. Healthcare professionals' perceived problems in fast-track hip and knee arthroplasty: results of a qualitative interview study. J Orthop Surg Res 2019 Sep 04;14(1):294 [FREE Full text] [doi: 10.1186/s13018-019-1334-3] [Medline: 31484536]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
mHealth: mobile health
MRI: magnetic resonance imaging
PDA: personal digital assistants
QR: quick response
SHG: serious health game

Edited by C Lovis; submitted 21.08.20; peer-reviewed by X Garcia-Eroles, M Lall; comments to author 06.10.20; revised version received 17.10.20; accepted 18.10.20; published 13.11.20.

Please cite as:
Rantala A, Jansson MM, Helve O, Lahdenne P, Pikkarainen M, Pölkki T
Parental Experiences of the Pediatric Day Surgery Pathway and the Needs for a Digital Gaming Solution: Qualitative Study
JMIR Med Inform 2020;8(11):e23626
URL: http://medinform.jmir.org/2020/11/e23626/
doi:10.2196/23626
PMID:33185556

Original Paper

Web- and Artificial Intelligence–Based Image Recognition For Sperm Motility Analysis: Verification Study

Vincent FS Tsai[1,2], MD, PhD; Bin Zhuang[3], MEng; Yuan-Hung Pong[2,4], MD; Ju-Ton Hsieh[2], MD; Hong-Chiang Chang[2], MD

[1]Department of Urology, Ten-Chan General Hospital, Taoyuan, Taiwan

[2]Department of Urology, National Taiwan University Hospital, Taipei, Taiwan

[3]Division of Research and Development, Createcare Technology Corporation, Shenzhen, China

[4]Department of Urology, Ten-Chen General Hospital, Taoyuan, Taiwan

Corresponding Author:
Hong-Chiang Chang, MD
Department of Urology
National Taiwan University Hospital
7, Zhung-Shan S. Road
Taipei, 100
Taiwan
Phone: 886 223123456 ext 62135
Email: bird8873@gmail.com

Abstract

Background: Human sperm quality fluctuates over time. Therefore, it is crucial for couples preparing for natural pregnancy to monitor sperm motility.

Objective: This study verified the performance of an artificial intelligence–based image recognition and cloud computing sperm motility testing system (Bemaner, Createcare) composed of microscope and microfluidic modules and designed to adapt to different types of smartphones.

Methods: Sperm videos were captured and uploaded to the cloud with an app. Analysis of sperm motility was performed by an artificial intelligence–based image recognition algorithm then results were displayed. According to the number of motile sperm in the vision field, 47 (deidentified) videos of sperm were scored using 6 grades (0-5) by a male-fertility expert with 10 years of experience. Pearson product-moment correlation was calculated between the grades and the results (concentration of total sperm, concentration of motile sperm, and motility percentage) computed by the system.

Results: Good correlation was demonstrated between the grades and results computed by the system for concentration of total sperm (r=0.65, *P*<.001), concentration of motile sperm (r=0.84, *P*<.001), and motility percentage (r=0.90, *P*<.001).

Conclusions: This smartphone-based sperm motility test (Bemaner) accurately measures motility-related parameters and could potentially be applied toward the following fields: male infertility detection, sperm quality test during preparation for pregnancy, and infertility treatment monitoring. With frequent at-home testing, more data can be collected to help make clinical decisions and to conduct epidemiological research.

(JMIR Med Inform 2020;8(11):e20031) doi:10.2196/20031

KEYWORDS

Male infertility; semen analysis; home sperm test; smartphone; artificial intelligence; cloud computing; telemedicine

Introduction

Infertility is a worldwide problem, with a prevalence of 15% [1], and sperm plays an important role [2]. In the process of fertilization, sperm are initially ejaculated around the cervix and then swim to the proximal oviduct, where they encounter the oocyte and accomplish fertilization [3]. To accomplish fertilization, large numbers of sperm are needed to overcome the filtration of cervical mucus while progressive motility is necessary for sperm to travel for a long distance. Human sperm concentration and motility fluctuate over time [4]. Therefore, frequent monitoring of sperm concentration and motility is crucial for couples who are preparing for spontaneous pregnancy.

Conventionally, men are asked to have semen analysis performed at a doctor's office or laboratory, a procedure that is both time-consuming and embarrassing for men, according to our clinical observations and prior literature [5]. Hence, it makes frequent measurement of sperm concentration and motility difficult. To tackle this problem, several trials of home sperm tests were developed in the past few decades, such as SpermCheck [6], Fertell [7], and Trak Male Fertility Testing System [8]. Some, such as YO sperm test and the Kobori single ball-lens system, attracted a lot of attention; both are systems that utilizes platforms based on smartphone systems [9,10]. There are some advantages to using smartphone-based home sperm tests, such as high-resolution video recording, robust calculation ability, and accessible internet communication. All these advantages can allow users to be tested in a setting that encourages more frequent measurements, while preserving the privacy and accuracy of results, similar to those of other point-of-care test systems that have been applied in measurements of blood sugar, blood pressure, and body temperature.

Bemaner (Shenzhen Createcare Technology Co) is a smartphone-based home sperm motility measurement system composed of a microscope and microfluidic modules and is designed to adapt to different types of smartphone designs and interfaces (Figure 1). Bemaner is quite different from the YO sperm test and the Kobori single ball-lens system [10]. The YO sperm test utilizes a tailor-made adapter (slide) to fit specific types of smartphones, but Bemaner can fit all smartphones currently. Sperm videos are captured and uploaded to the cloud from the smartphone via software apps. Unlike the Kobori single ball-lens system, in which motile and static sperm are counted by a person [10], for Bemaner, analysis of sperm motility is performed by an artificial intelligence (AI) image recognition algorithm, and results are displayed to end users on the smartphone interface. The purpose of this study was to verify the performance of the AI sperm image recognition algorithm and cloud computing system.

Figure 1. Bemaner smartphone-based home sperm motility test system.

Methods

Video Clips of Motile Sperm and Results Computed by AI

Semen samples can be collected at home by users. After 30 minutes of liquefaction, the semen sample is dipped by a small biochip cup (analog to glass slide) and then covered by a large biochip cup (analog to cover slip). The space containing semen samples between these two cups is designed to be 10 micrometers deep, which can contain a single layer of sperm for a specific volume of 0.2 microliters. Through the microscopic and microfluidic modules of the device, the video clips of motile sperm can be captured and uploaded by any recent smartphone, which is easily aligned to the microscopic and microfluidic modules (Multimedia Appendix 1). Deidentified (ie, users' information removed) video clips of motile sperm were retrieved from the central cloud computing server (Alibaba cloud, Hangzhou, China). With the process of deidentification, it was impossible to obtain specific user consent for this study. However, the security and privacy of users'

information were well protected through the process of deidentification.

The results, including the concentration of total sperm, the concentration of motile sperm, and the motility percentage, computed by the AI image recognition algorithm (version 1.0.8_5/22) with cloud computing on the central server were also retrieved with every video clip of motile sperm. In detail, concentration of total sperm and motile sperm were derived by the image recognition AI algorithm. Motility percentage is calculated as the quotient of concentration of motile to total sperm.

Scoring of Motile Sperm

The gold standard for assessing semen quality in the current World Health Organization manual [4] is analysis performed under a microscope by a well-trained professional staff. According to the number of motile sperm in the vision field of each video clip, the 47 deidentified videos of motile sperm were classified into 6 grades (0-5) by a male-fertility expert with 10 years of experience. The criteria were the following: grade 0, there are no areas of motile sperm in the field of vision; grade 1, some motile sperm; grade 2; the amount of motile sperm is between grade 1 and 3; grade 3, motile sperm occupy half of the field of vision; grade 4, the number of motile sperm is between grade 3 and 5; grade 5, motile sperm occupy almost the whole field of vision (Multimedia Appendix 2-Multimedia Appendix 7).

Relationship Between Human and AI Results

The grade assessed by a male-fertility expert was an ordinal variable; the relationships between grade and concentration of total sperm, concentration of motile sperm, and motility percentage was calculated determined with Pearson product-moment correlation coefficients. Analyses were calculated using the software Excel (Microsoft Inc, 2013). A *P* value less than .01 was considered statistically significant.

Results

The results (concentration of total sperm, concentration of motile sperm and motility percentage) of the AI algorithm and the distribution of scored grade according to the number of motile sperm in the vision field of each video clip are is shown in Table 1.

Relationships between the grades and Bemaner AI algorithm results were r=0.65 (P<.001) for concentration of total sperm (Figure 2), r=0.90 (P<.001) for motility percentage (Figure 3), and r=0.84 (P<.001) for concentration of motile sperm (Figure 4).

Table 1. Results.

Source and variable	Value
Bemaner AI algorithm-based, mean (SD)	
Concentration of total sperm (million cell/mL)	111.02 (72.13)
Concentration of motile sperm (million cell/mL)	50.87 (61.47)
Motility percentage (%)	32.8 (32.5)
Sperm motility count by expert grade (million cell/mL), n	
Grade 0	7
Grade 1	8
Grade 2	8
Grade 3	5
Grade 4	14
Grade 5	5

Figure 2. Relationship between Bemaner concentration of total sperm and scored grades.

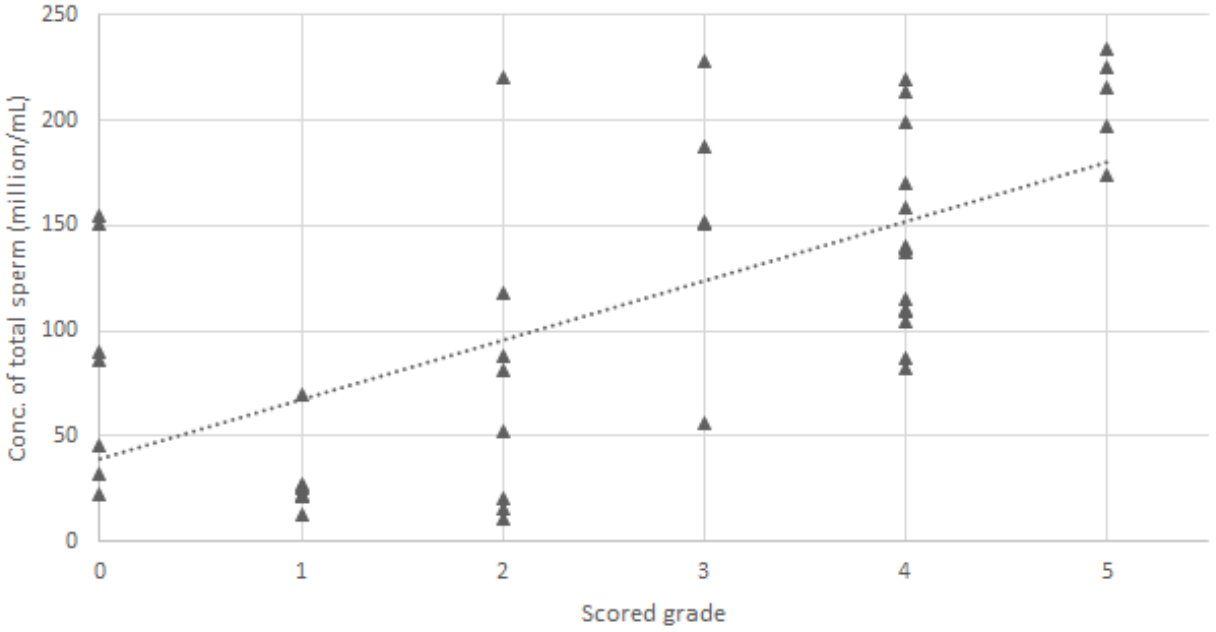


Figure 3. Relationships between Bemaner motility percentage and scored grades.

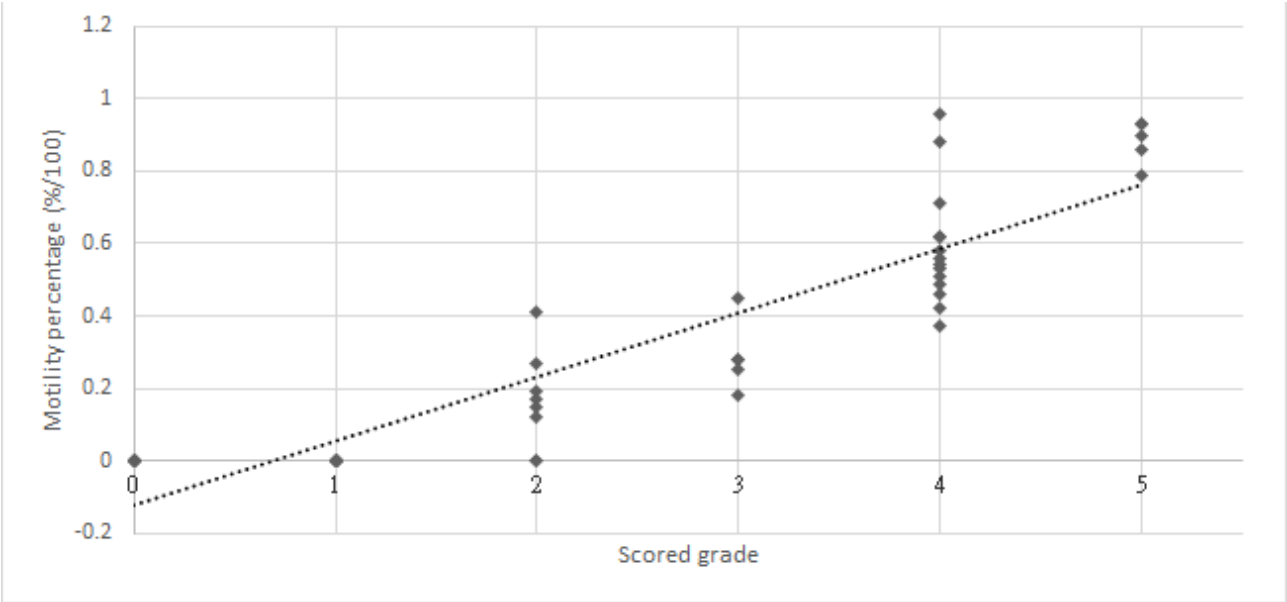
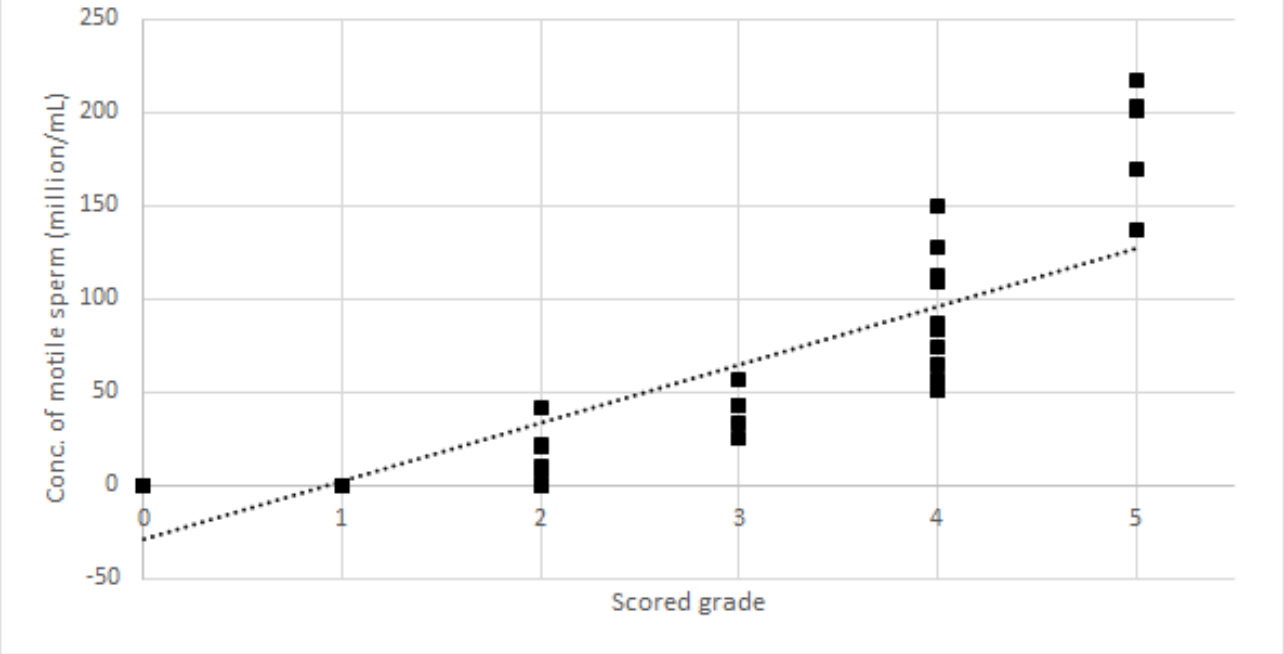


Figure 4. Relationships between Bemaner concentration of motile sperm and scored grades.



Discussion

To our knowledge, this is the first smartphone-based system for semen analysis accomplished through an AI image recognition algorithm on cloud computing. Even though there are other smartphone-based systems, such as YO sperm test [9] and Kobori single ball-lens system [10], neither utilizes an AI imaging recognition algorithm with cloud computing. YO sperm test analyzes sperm videos by an app installed on a smartphone, while Kobori single ball-lens system projects the mobile phone captured videos onto a desktop monitor and counts sperm parameters manually.

Additionally, there are some benefits for data processing through cloud computing. Only requiring connection with the internet, the system provides an increased flexibility and possibility in terms of setting up test locations and test frequencies. It is also easier for engineers to perform system maintenance such as algorithm revision and software updates instead of frequent update requests on the user end smartphone app. Also, accumulated big data may demonstrate important epidemiological characteristics of male fertility (eg, regional or temporal). Such data processes, based on cloud computing, are quite novel and promising for large-scaled epidemiological studies and public-health related policy making.

The results calculated by Bemaner and grades assessed by a male-fertility expert, motility percentage and concentration of motile sperm, demonstrated better correlations with the scored grade. The respective correlation coefficients are 0.90 (P<.001) and 0.84 (P<.001). This means that motility percentage and concentration of motile sperm calculated by Bemaner were comparable to a male-fertility expert's judgment for assessing sperm motility.

However, concentration of total sperm was only moderately correlated (r=0.65; P<.001) with the scored grade. This is attributed to some of total sperm being immobile sperm, which cannot achieve natural fertilization [11,12].

With regard to corresponding motile sperm concentration of each scored grade, the calibrated reference values of sperm concentration are provided (Table 2).

Table 2. Calibrated reference values of sperm concentration.

Grade	Range (million cells/mL)
0	0
1	<13.9
2	0-27.7
3	28.3-51.9
4	50.7-113.9
5	153.4-217.6

The accuracy of these values can be improved as the accumulated data size grows. This is also one of the benefits that these types of online processes for point-of-care test results can provide.

In comparison with other smartphone-based sperm analysis systems, such as YO sperm test [9] and the Kobori single ball-lens system [10], Bemaner can provide users with information about sperm concentration, motile sperm

percentage, and motile sperm concentration. These parameters are essential for a man to achieve natural fertilization. That may be the reason why these parameters are chosen for screening tests [13,14].

Graded results show more information about fluctuation of sperm concentration and motility than results indicating positive or negative, such as SpermCheck [6] and Fertell [7]. It is also more easily understood by lay users. YO sperm test and Bemaner, in the latest version of app on the smartphone provides not only total sperm concentration, motile sperm concentration, and percentage of motile sperm but also graded results of the motile sperm concentration based on the findings of this study. Both provide graded results for motile sperm concentration (ie, YO score and Bemaner scored grade). It can be used as a home sperm test instead of having comprehensive and robust sperm parameters assessments in a fertility laboratory [4]. Practically, graded results can be applied in some scenarios that require information about fluctuating test results, such as male infertility detection, sperm quality test during preparation for pregnancy, and infertility treatment monitoring.

YO sperm test obtained test results by the installed algorithm and the results were compared to those of another image-recognition sperm test SQA-Vision (Medical Electronics Systems) by Agarwal et al [9]. Kobori et al [10] counted sperm on images projected to desktop screen and compared their results with computer assisted semen analysis. In contrast, Bemaner used an AI algorithm with cloud computing to analyze images for sperm testing. The results were compared with the grades assessed by an experienced male-fertility expert. Because the images for analysis were uploaded by end users and deidentified, it was impossible to acquire any original parameters of the corresponding semen samples. (The appearance, viscosity, pH, total sperm concentration and motile sperm concentration manually measured under microscope for a semen sample are original parameters of the sample. Only the previous captured and uploaded videos of sperm could be retrieved in this study.) Therefore, grades of sperm images were utilized as the reference method. This is one of the limitations of the study.

Currently, the aforementioned 3 systems cannot analyze sperm images to assess sperm morphology according to the World Health Organization morphology assessment paradigm [4]. A new version of the AI algorithm including morphology assessment is currently in development for the Bemaner system.

In the face of emerging pandemics causing city lockdown and health care need for remote regions, tele-medicine will prevail and requires on-line processes of data and analysis connected with off-line point-of-care tests for end users. The Bemaner system is a prime example of this type of implementation, especially for infertility, which is always related to the following aspects: environmental pollution, life style, and diet [15-17], the analysis of big data could contribute much more than traditional case reports in research [18].

The framework of data processing in this system can be reconstructed as the big data grows and we extend our scope to collect the different aspects related to users' life style, diet, environmental pollutions, and medical interventions. In the future, some techniques of cloud computing and data storage to optimize big data processing could be applied to help speed the process, such as adopting concepts such as intermediate data caching. Nevertheless, the current process of data transmission and analyzing is similar to an application program interface call, in which a request with the data is sent to the server, the server processes the data with our image recognition algorithm to generate the result, and the server responses with an answer.

There are still some improvements to be implemented in the near future. The biochip cups for containing semen specimens could be adjusted automatically by a step-motor to increase the scope of observation and the number of observed sperm. The optic module could be modified to increase the magnification to define sperm's morphology according to the WHO criteria [4]. Revisions of the AI image recognition algorithm will continue to improve as data accumulates, enhancing its accuracy. Big data for this system has huge potential to introduce various insight, both in clinical and epidemiological fields, to improve human fertility.

This smartphone-based system for measuring sperm motility (Bemaner) accurately measures parameters related to sperm motility and could potentially be applied toward the following fields: male infertility detection, sperm quality test during preparation for pregnancy, and infertility treatment monitoring. With frequent testing, more data can be collected to help make clinical decisions and conduct epidemiological studies.

The overall contributions of this paper are (1) to introduce a novel online system for semen analysis calculated by an AI image recognition algorithm combined with cloud computing technology; (2) to prove the performance of the system by correlating human and AI results; (3) to suggest potential applications in male infertility detection, sperm quality test during preparation for pregnancy, and infertility treatment monitoring; and (4) to foresee that the big data collected by this system can play an important role in making clinical decisions and conducting epidemiological research.

Acknowledgments

Authors want to thank Yu-Chian Tsai from the University of Michigan for his ideas about cloud computing and device networking.

Authors' Contributions

H-CC and VF-ST contributed to the study concept and design. BZ contributed to the technology and devices implementation. Y-HP contributed to acquisition of data. J-TH contributed to the statistical analysis. H-CC and VF-ST contributed to the interpretation of results. J-TH and VF-ST drafted the manuscript. All authors gave final approval for the manuscript.

Conflicts of Interest

BZ is the chief of research and development at Createcare. VF-ST is a medical consultant for Createcare.

Multimedia Appendix 1
Operating process of Bemaner.
[MOV File , 168101 KB - medinform_v8i11e20031_app1.mov]

Multimedia Appendix 2
Scored grades of sperm videos: Grade 0.
[MOV File , 9520 KB - medinform_v8i11e20031_app2.mov]

Multimedia Appendix 3
Scored grades of sperm videos: Grade 1.
[MOV File , 12746 KB - medinform_v8i11e20031_app3.mov]

Multimedia Appendix 4
Scored grades of sperm videos: Grade 2.
[MOV File , 12138 KB - medinform_v8i11e20031_app4.mov]

Multimedia Appendix 5
Scored grades of sperm videos: Grade 3.
[MOV File , 20759 KB - medinform_v8i11e20031_app5.mov]

Multimedia Appendix 6
Scored grades of sperm videos: Grade 4.
[MOV File , 19378 KB - medinform_v8i11e20031_app6.mov]

Multimedia Appendix 7
Scored grades of sperm videos: Grade 5.
[MOV File , 21784 KB - medinform_v8i11e20031_app7.mov]

References

1. Agarwal A, Mulgund A, Hamada A, Chyatte MR. A unique view on male infertility around the globe. Reprod Biol Endocrinol 2015 Apr 26;13:37 [FREE Full text] [doi: 10.1186/s12958-015-0032-1] [Medline: 25928197]
2. Nosrati R, Graham PJ, Zhang B, Riordon J, Lagunov A, Hannam TG, et al. Microfluidics for sperm analysis and selection. Nat Rev Urol 2017 Dec;14(12):707-730. [doi: 10.1038/nrurol.2017.175] [Medline: 29089604]
3. Suarez S, Pacey A. Sperm transport in the female reproductive tract. Hum Reprod Update 2006;12(1):23-37. [doi: 10.1093/humupd/dmi047] [Medline: 16272225]
4. Laboratory manual for the examination and processing of human semen. 5th edition. Geneva: World Health Organization; 2010.
5. Elzanaty S, Malm J. Comparison of semen parameters in samples collected by masturbation at a clinic and at home. Fertil Steril 2008 Jun;89(6):1718-1722. [doi: 10.1016/j.fertnstert.2007.05.044] [Medline: 17658521]
6. Coppola M, Klotz K, Kim K, Cho H, Kang J, Shetty J, et al. SpermCheck Fertility, an immunodiagnostic home test that detects normozoospermia and severe oligozoospermia. Hum Reprod 2010 Apr;25(4):853-861 [FREE Full text] [doi: 10.1093/humrep/dep413] [Medline: 20139122]
7. Björndahl L, Kirkman-Brown J, Hart G, Rattle S, Barratt C. Development of a novel home sperm test. Hum Reprod 2006;21(1):145-149. [doi: 10.1093/humrep/dei330] [Medline: 16267078]
8. Schaff UY, Fredriksen LL, Epperson JG, Quebral TR, Naab S, Sarno MJ, et al. Novel centrifugal technology for measuring sperm concentration in the home. Fertil Steril 2017 Feb;107(2):358-364.e4. [doi: 10.1016/j.fertnstert.2016.10.025] [Medline: 27887718]
9. Agarwal A, Panner Selvam MK, Sharma R, Master K, Sharma A, Gupta S, et al. Home sperm testing device versus laboratory sperm quality analyzer: comparison of motile sperm concentration. Fertil Steril 2018 Dec;110(7):1277-1284. [doi: 10.1016/j.fertnstert.2018.08.049] [Medline: 30424879]
10. Kobori Y, Pfanner P, Prins GS, Niederberger C. Novel device for male infertility screening with single-ball lens microscope and smartphone. Fertil Steril 2016 Sep 01;106(3):574-578. [doi: 10.1016/j.fertnstert.2016.05.027] [Medline: 27336208]

11. Ombelet W, Dhont N, Thijssen A, Bosmans E, Kruger T. Semen quality and prediction of IUI success in male subfertility: a systematic review. Reprod Biomed Online 2014 Mar;28(3):300-309. [doi: 10.1016/j.rbmo.2013.10.023] [Medline: 24456701]
12. Wang C, Swerdloff RS. Limitations of semen analysis as a test of male fertility and anticipated needs from newer tests. Fertil Steril 2014 Dec;102(6):1502-1507 [FREE Full text] [doi: 10.1016/j.fertnstert.2014.10.021] [Medline: 25458617]
13. Buck Louis GM, Sundaram R, Schisterman EF, Sweeney A, Lynch CD, Kim S, et al. Semen quality and time to pregnancy: the Longitudinal Investigation of Fertility and the Environment Study. Fertil Steril 2014 Feb;101(2):453-462 [FREE Full text] [doi: 10.1016/j.fertnstert.2013.10.022] [Medline: 24239161]
14. Almeida S, Rato L, Sousa M, Alves MG, Oliveira PF. Fertility and sperm quality in the aging male. Curr Pharm Des 2017 Nov 28;23(30):4429-4437. [doi: 10.2174/1381612823666170503150313] [Medline: 28472913]
15. Nassan FL, Jensen TK, Priskorn L, Halldorsson TI, Chavarro JE, Jørgensen N. Association of dietary patterns with testicular function in young Danish men. JAMA Netw Open 2020 Feb 05;3(2):e1921610 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.21610] [Medline: 32083688]
16. Kasman AM, Del Giudice F, Eisenberg ML. New insights to guide patient care: the bidirectional relationship between male infertility and male health. Fertil Steril 2020 Mar;113(3):469-477. [doi: 10.1016/j.fertnstert.2020.01.002] [Medline: 32089256]
17. Poli D, Andreoli R, Moscato L, Pelà G, de Palma G, Cavallo D, et al. The relationship between widespread pollution exposure and oxidized products of nucleic acids in seminal plasma and urine in males attending a fertility center. Int J Environ Res Public Health 2020 Mar 13;17(6):1880 [FREE Full text] [doi: 10.3390/ijerph17061880] [Medline: 32183208]
18. Patel DP, Jenkins TG, Aston KI, Guo J, Pastuszak AW, Hanson HA, et al. Harnessing the full potential of reproductive genetics and epigenetics for male infertility in the era of "big data". Fertil Steril 2020 Mar;113(3):478-488 [FREE Full text] [doi: 10.1016/j.fertnstert.2020.01.001] [Medline: 32089255]

Abbreviations

AI: artificial intelligence
WHO: World Health Organization

Edited by G Eysenbach; submitted 18.05.20; peer-reviewed by Z Yang, F Palmieri; comments to author 30.06.20; revised version received 11.07.20; accepted 28.10.20; published 19.11.20.

Please cite as:
Tsai VFS, Zhuang B, Pong YH, Hsieh JT, Chang HC
Web- and Artificial Intelligence–Based Image Recognition For Sperm Motility Analysis: Verification Study
JMIR Med Inform 2020;8(11):e20031
URL: http://medinform.jmir.org/2020/11/e20031/
doi:10.2196/20031
PMID:33211025

Original Paper

Patient Triage by Topic Modeling of Referral Letters: Feasibility Study

Irena Spasic[1], PhD; Kate Button[2], PhD

[1]School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom

[2]School of Healthcare Sciences, Cardiff University, Cardiff, United Kingdom

Corresponding Author:
Irena Spasic, PhD
School of Computer Science & Informatics
Cardiff University
5 The Parade
Cardiff, CF24 3AA
United Kingdom
Phone: 44 02920870320
Email: spasici@cardiff.ac.uk

Abstract

Background: Musculoskeletal conditions are managed within primary care, but patients can be referred to secondary care if a specialist opinion is required. The ever-increasing demand for health care resources emphasizes the need to streamline care pathways with the ultimate aim of ensuring that patients receive timely and optimal care. Information contained in referral letters underpins the referral decision-making process but is yet to be explored systematically for the purposes of treatment prioritization for musculoskeletal conditions.

Objective: This study aims to explore the feasibility of using natural language processing and machine learning to automate the triage of patients with musculoskeletal conditions by analyzing information from referral letters. Specifically, we aim to determine whether referral letters can be automatically assorted into latent topics that are clinically relevant, that is, considered relevant when prescribing treatments. Here, clinical relevance is assessed by posing 2 research questions. Can latent topics be used to automatically predict treatment? Can clinicians interpret latent topics as cohorts of patients who share common characteristics or experiences such as medical history, demographics, and possible treatments?

Methods: We used latent Dirichlet allocation to model each referral letter as a finite mixture over an underlying set of topics and model each topic as an infinite mixture over an underlying set of topic probabilities. The topic model was evaluated in the context of automating patient triage. Given a set of treatment outcomes, a binary classifier was trained for each outcome using previously extracted topics as the input features of the machine learning algorithm. In addition, a qualitative evaluation was performed to assess the human interpretability of topics.

Results: The prediction accuracy of binary classifiers outperformed the stratified random classifier by a large margin, indicating that topic modeling could be used to predict the treatment, thus effectively supporting patient triage. The qualitative evaluation confirmed the high clinical interpretability of the topic model.

Conclusions: The results established the feasibility of using natural language processing and machine learning to automate triage of patients with knee or hip pain by analyzing information from their referral letters.

(JMIR Med Inform 2020;8(11):e21252) doi:10.2196/21252

KEYWORDS

natural language processing; machine learning; data science; medical informatics; computer-assisted decision making

Introduction

Background

Currently, a pathway recommended for musculoskeletal conditions such as knee or hip pain consists of their management within primary care followed by referral to a multiprofessional assessment and treatment clinic if a specialist opinion is required [1]. The aging population increases the demand for health care resources [2], emphasizing the need to streamline care pathways to maximize efficiency and ensure patients receive optimal care

for their needs. With this aim, referral prioritization systems were developed for hip and knee pain and tested to fast-track cases for surgical opinion based on referral information provided by the primary care [3,4]. However, their prioritization criteria lacked adequate sensitivity and specificity for patients moving between surgical and conservative pathways. Information conveyed in referral letters underpins the referral decision-making process, but it has not been explored systematically for the purposes of treatment prioritization for musculoskeletal conditions. Automated analysis of referral letters can identify variables that can be used alongside demographic and health-related data to improve treatment prioritization. Within the context of musculoskeletal conditions, natural language processing (NLP) was used successfully to automate the analysis of radiology reports [5,6] and patient questionnaires [7].

Indeed, NLP has repeatedly demonstrated its feasibility to extract clinical variables from clinical narratives, making them available for large-scale analysis down the stream [8]. Traditionally, rule-based approaches have been commonly used to extract variables of predefined types [9]. Machine learning has long been hailed as a silver bullet solution for the knowledge elicitation bottleneck, the main argument being that the task of annotating the data manually is easier than that of eliciting the knowledge. However, a recent systematic review of machine learning approaches based on clinical text data revealed the data annotation bottleneck to be one of the key obstacles to machine learning approaches in clinical NLP [10]. However, the biggest challenge for these applications to become part of routine clinical practice is the problem of human interpretability of automated outputs. Machine learning approaches may offer faster development of algorithms and their performance improvement, but some do so at the expense of the interpretability of the results [11]. Topic modeling can kill both birds with one stone. First, the aim of topic modeling is to identify latent topics that can be used to organize a corpus, where each document contains a mixture of topics in different proportions. As an unsupervised method, it does not require data to be annotated manually. This means that the algorithm can readily utilize vast amounts of data, allowing the machine learning model to more accurately capture statistically significant patterns. Second, each topic is associated with a set of words that are extracted automatically from the corpus based on their distribution. The highest-ranked words can help interpret the underlying semantics.

Related Work

A popular topic modeling algorithm is the latent Dirichlet allocation (LDA) [12]. LDA is a three-level hierarchical Bayesian model in which each document is modeled as a finite mixture over an underlying set of topics and each topic is modeled as an infinite mixture over an underlying set of topic probabilities. Although LDA is used frequently in NLP research, it is yet to make a significant mark on clinical NLP, which is still heavily biased in favor of supervised learning methods [10]. Nonetheless, LDA is steadily finding its clinical applications, such as improving clinical process efficiency [13-15], predicting hospital readmission [16], patient safety [17-19], and patient phenotyping [20-22]. Some of the topic models were specifically evaluated for interpretability from a clinician's perspective [14,16]. To improve coherence and interpretability of topics, some approaches combined LDA with clinical terminologies, such as the Medical Dictionary for Regulatory Activities [18] and the Systematized Nomenclature of Medicine Clinical Terms [15]. Typical reasons cited for choosing LDA over supervised learning approaches include alleviating the need for labor-intensive data annotation, avoiding human annotation bias, and the potential to identify latent topics in the data that may not be apparent a priori. The latter is particularly important in clinical scenarios with *unknown unknowns*, such as patient safety [17-19]. In terms of training a topic model, many approaches struggled to fine-tune the number of topics as one of the key hyperparameters of the LDA algorithm. In most cases, a plausible justification for the number of topics was lacking, for example, 25 [20], 100 [17,18], 75 [16], 50/100/150 [14], and 50/100/200 [21].

The research gaps identified in this overview of related work are as follows. Despite finding various clinical applications, LDA is yet to be used to support triage. The biggest challenge for these applications to become widely adopted in clinical practice is the perception of interpretability. However, few studies have specifically evaluated the interpretability of the LDA outputs from a clinician's perspective. Clinical terminologies have been combined with the LDA to improve interpretability, but the resources used to support such functionality do not include the Unified Medical Language System (UMLS), which offers a unique opportunity to abstract clinical concepts into higher categories of knowledge. Finally, for the topics to be easily distinguishable (and, hence, interpretable), their number needs to reflect the latent themes and patterns present in a given data set. However, none of the considered approaches provided a strategy to infer the value of this hyperparameter from the data. In this study, we addressed these four gaps.

First, we applied the LDA to a corpus of referral letters and used topics as features to automatically classify each letter against a list of potential treatments. This can then be used to automate patient triage, that is, assort them into priority groups according to their medical needs. Second, we proposed a novel method for evaluating the interpretability of topics. Third, we used the UMLS to incorporate the interpretation of clinical concepts at different levels of abstraction into the LDA. Finally, we systematically fine-tuned the number of topics using a measure of topic coherence.

Methods

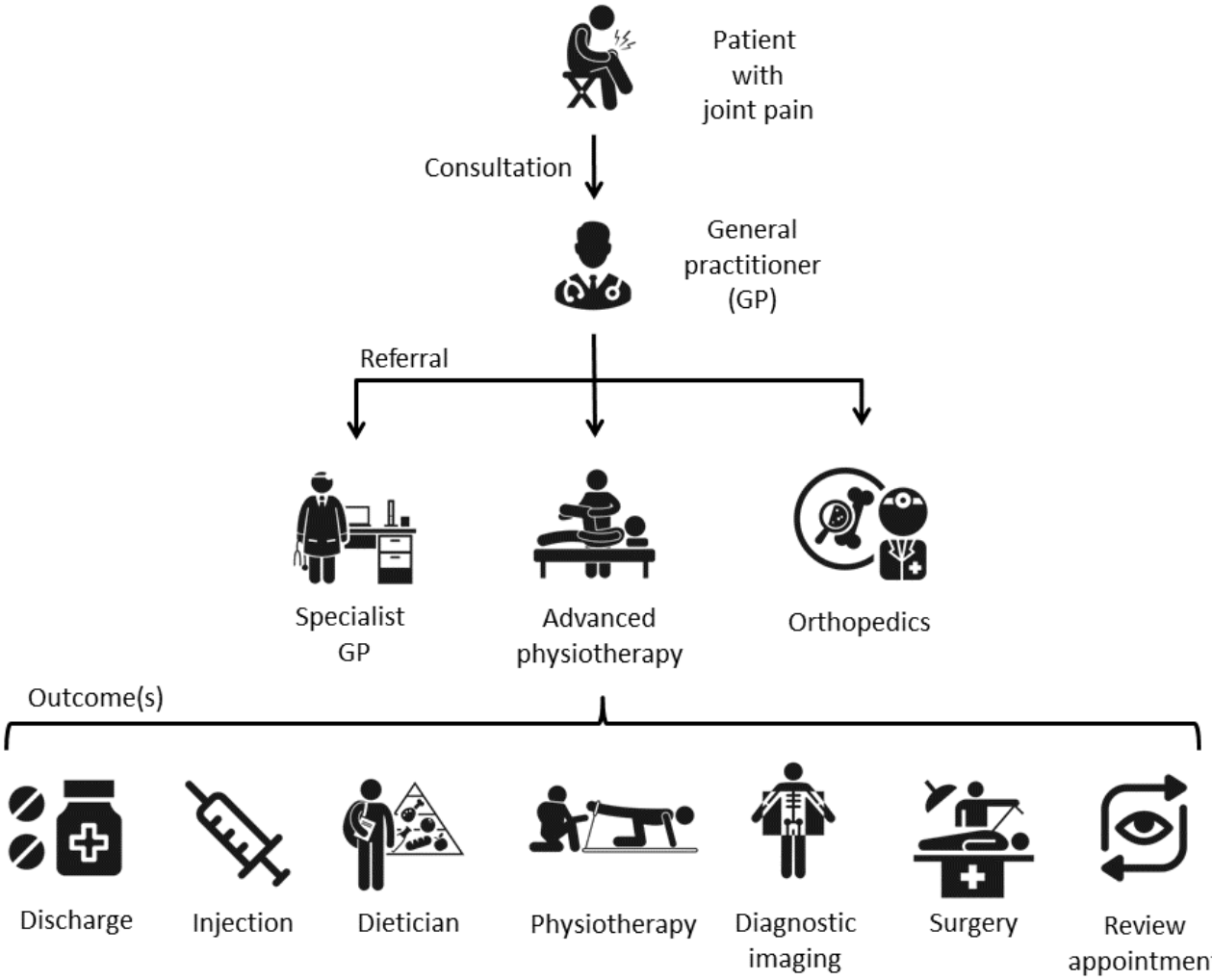
Data Collection

Data collection was originally described in the study by Button et al [23]. In summary, patients were eligible to take part in the study if they were referred by their general practitioner for joint (knee or hip) pain, they were aged 18 years or older, they could provide informed consent, and they could speak English fluently. The exclusion criteria included pain secondary to other health conditions such as rheumatoid arthritis, pain secondary to joint replacement, surgery for the same joint within the last 12 months, or having already received treatment at the

primary-secondary care interface for the same condition within the last 6 months.

The care pathway is illustrated in Figure 1. A patient with joint pain is referred by a clinician from their general practice to a specialist clinic in secondary care, which could be an orthopedic clinic, general practice with musculoskeletal specialism, or advanced physiotherapy clinic. Appropriate treatment is suggested when the patient is seen in secondary care.

Figure 1. Musculoskeletal care pathway for adults with hip and knee pain. GP: general practitioner.



Patients were recruited from one Local Health Board, an administrative unit within the National Health Service in Wales, which supports a population of around 445,000 people. A total of 634 participants were recruited between August 2016 and January 2017, and their referral letters were collected. The follow-up data collection was completed in June 2018. This included recording of any treatments performed. A subset of 576 patients with complete data, including the original referral letter and the corresponding treatments, was used in this study. The distribution of their treatments is given in Table 1. Note that a single patient may have had multiple treatments.

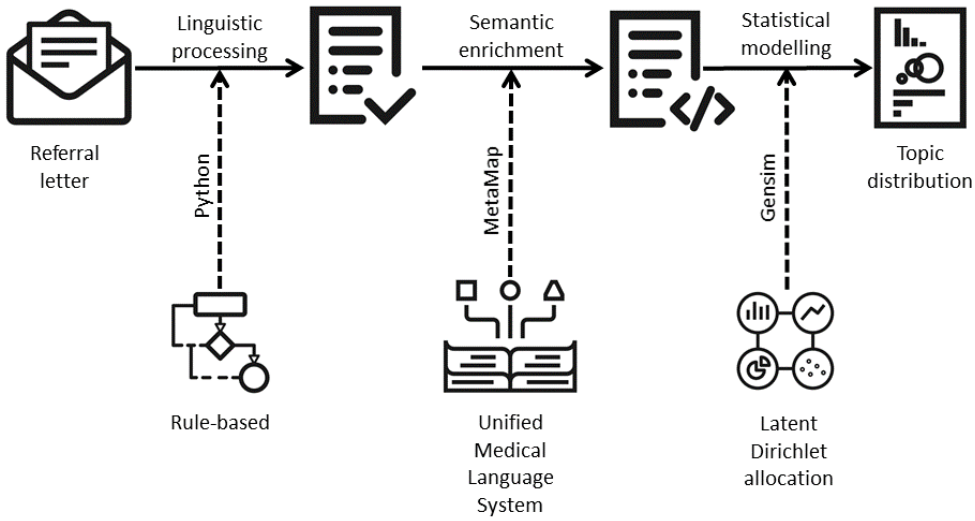
Table 1. The distribution of treatment referrals.

ID	Treatment	Total number of patients, n
O1	Orthopedic referral	53
O2	Discharge (no further appointments booked)	173
O3	Injection	101
O4	Nutritionist	15
O5	Physiotherapy	152
O6	Diagnostic imaging	112
O7	Surgery	99
O8	Review appointment	223
O9	Any other referral	16

System Design

The main research question addressed in this study is as follows: Can triaging patients (into cohorts) based on their referral letters be semiautomated? To that end, we designed a system that can support referral decision making (Figure 2). A corpus of referral letters was used to train a topic model with the ultimate aim of using topics to narrow down the choice of potential treatments and streamline the referral pathway. To reduce potential overfitting to a relatively small training data set, we regularized and generalized its text content. First, the text was regularized by applying a set of linguistic rules designed to reduce idiosyncrasies associated with clinical sublanguage, covering punctuation, acronyms, abbreviations, orthographic and lexical variation, and personal names of patients and clinicians. Subsequently, an external medical language system was used to effectively normalize the terminology used, making the topic model robust with respect to terminological variation. The following sections describe the three modules in greater detail.

Figure 2. System design for topic modeling of referral letters.



Linguistic Processing

The linguistic preprocessing and normalization module originally developed to support cohort selection from hospital discharge summaries was adapted for this study [24]. In addition to standard linguistic preprocessing operations, this module also handles punctuation in clinical narratives, which can affect the results of text segmentation algorithms developed for general language [25]. However, its main purpose is to streamline subsequent text analysis and reduce overfitting by regularizing the text content. This involves basic string operations such as lowercasing, fully expanding enclitics, and special characters. It further normalizes text content by replacing a selected subset of words and phrases with their representatives. Here, special consideration is given to acronyms and abbreviations as they are known to have a major impact on the retrieval of relevant information [26]. These mappings are supported by a set of local lexica whose content was adapted for this study to support migration from the domain of hospital discharge summaries to that of referral letters. To facilitate this process, we extracted multiword terms (including their acronyms) from referral letters automatically using FlexiTerm [27,28] and manually curated the list of conflated term variants.

New functionality added to the linguistic processing module includes recognition of personal names. Personal names, like any other words, can be selected automatically as topic descriptors. For example, if several patients were referred to Dr Jane Doe, who is a physiotherapist, then her name may become correlated with a *physiotherapy theme* in referral letters,

ultimately resulting in the words “Jane” and “Doe” emerging as the topic descriptors. Not only are these words not informative of the topic but they also cannot be generalized to other data sets where these names do not exist, or they refer to different persons, thus rendering the model either inapplicable or inaccurate. To prevent a topic model from overfitting to personal names, they are replaced by a generic representative. For this purpose, we originally considered existing named entity recognition libraries (eg, [29,30]) to recognize personal names in referral letters. However, having been designed with general language in mind, their overzealous matching algorithm could not distinguish between different uses of personal names. As illustrated by the taxonomy for the rehabilitation of knee conditions [31], many clinically relevant concepts feature personal names, for example, Hoffa fat pad, Baker cyst, or McMurray test. Replacing these mentions of personal names with generic representatives would remove important content that can be used to describe a topic. On the other hand, referral letters are written using a formal style, which prescribes the use of honorifics. This fact was exploited to define a set of regular expressions based on honorifics and capitalization of personal names to automatically recognize the names of patients and clinicians. These names were replaced with a generic representative. This approach preserved personal names used to name body parts, diseases, tests, and any other medical concepts.

Semantic Enrichment

As a statistical model, a topic model may benefit from aggregating the distribution of synonyms (eg, “physio” and “physiotherapy”). Linking synonyms gives the model a better chance of capturing the semantics of underlying topics. Linguistic preprocessing implements lexical normalization, where both formal and informal abbreviations are translated to a standard vocabulary. For instance, “TKR” and “physio” would be translated to “total knee replacement” and “physiotherapy,” respectively. However, the problem of term variation may still persist. Examples from our corpus are many: “tear” versus “rupture,” “painkiller” versus “analgesic,” “oedema” versus “swelling,” “patella” versus “kneecap,” etc. The UMLS [32], which integrates multiple terminologies, classifications, and coding standards, maps such terms to concepts, which are assigned a concept unique identifier (CUI). A CUI can be used to markup synonymous terms in the text. Consider, for example, the sentences given in Textbox 1. Concept markups can be processed by topic modeling software similar to any other tokens in the corpus and, therefore, can be used as potential topic descriptors.

Textbox 1. Concept markups.

1. She struggles to take any *painkillers/C0002771* stronger than paracetamol.
2. He is opposed to regular *analgesics/C0002771*.
3. His recent magnetic resonance imaging shows *oedema/C0013604* and bursitis.
4. There is a little bit of *swelling/C0013604* of the knee joint.
5. The magnetic resonance imaging showed a complex *tear/C3203359* of the medial meniscus.
6. She has had a likely anterior cruciate ligament *rupture/C3203359*.

Moreover, concept markup can be used to effectively group together multiword expressions. This may improve the interpretability of topics. For example, when words describing a topic are presented independently of one another, such as “medial,” “joint,” “line,” and “tenderness” instead of “medial joint line tenderness,” then it is unclear whether the word “medial” refers to “meniscus” (“medial meniscus”), “ligament” (“medial collateral ligament”), “condyle” (“medial femoral condyle”) or indeed a “joint line” (“medial joint line”). Similarly, it remains unclear which anatomical entity is affected by “tenderness.” To alleviate this problem, topic modeling approaches often use an *n*-gram language model [33], with *n* being fixed to 2 and 3. Examples from our corpus (Textbox 2) illustrate that an *n*-gram approach may be too rigid for biomedical sublanguage, which is known for its terminological variability [27,28].

Textbox 2. Markup of multiword terms.

1. I could not reproduce pain with *McMurray test/C3669149*.
2. She does however experience pain on *McMurray* and Ege *testing/C3669149*.
3. He would be keen to consider a *total knee replacement/C0086511* as his pain has increased.
4. She is relatively young for consideration of *knee arthroplasty/C0086511*.
5. She has poor mobility following a few revisions of a right *knee prosthesis/C0086511*.
6. He is a 67-year-old male who has had *bilateral knee pain/C2220048* for a number of years.
7. She has persistent *pain in both knees/C2220048* with regular effusions.
8. She has crepitus in his left knee with *medial joint line tenderness/C0576135*.
9. No swelling of the knee but *tender medial joint line/C0576135*.
10. He had an effusion present and was *tender across his medial joint line/C0576135*.
11. On examination there was *tenderness along the joint line medially/C0576135*.

MetaMap, a highly configurable dictionary lookup software, can be used to discover the UMLS concepts in the text [34]. We used MetaMap to markup concepts such as those presented in Textboxes 1 and 2. Table 2 provides the most relevant details of the MetaMap configuration used. MetaMap also maps concepts to semantic types. Like CUIs, they can be used for markup. Semantic type markups can be used to unify concepts depicting a common theme. As examples from our corpus illustrate (Textbox 3), references to sports activities are very diverse. Individually, they may not be selected as topic descriptors because their occurrences are relatively rare. However, when they are mapped to their semantic type (*daily or recreational activity* (DORA)), we can observe common themes emerging focusing on age, fitness, and injury: young, physically active patients with a sports-related injury. These factors play an important role in recommending the most appropriate treatments. Their association with the given semantic type means that it could be a useful topic descriptor. For example, a clinician can reasonably assume that the given topic refers to a cohort of young, fit patients with a sports-related injury. Semantic type markups can be processed by topic modeling software similar to any other tokens in the corpus and, therefore, can be used as potential topic descriptors.

Table 2. MetaMap configuration.

Parameter	Description	Used	Rationale
a	Allows matching of acronyms and abbreviations.	No	These are the least reliable form of variation, for example, "OA" has got at least three full forms, for example, "osteoarthritis," "optic atrophy," and "ocular albinism." Local lexica were used in linguistic processing module instead to enforce tighter control of acronyms and abbreviations.
i	Ignores word order when matching a text phrase to a candidate concept name.	Yes	This option allows for syntactic variants such as "meniscus tear" and "tear of meniscus" to be conflated.
D	Forces the use of all derivational variants instead of only those between adjectives and nouns.	Yes	This option adds flexibility to conflation of syntactic variants such as torn/VBN meniscus/NN and meniscal/JJ tear/NN.
l	Enables retrieval of candidates for two-character words occurring in more than 2000 UMLS[a] strings and one-character words occurring in more than 1000 UMLS strings.	No	Like acronyms and abbreviations, short words are highly ambiguous.
8	Generates variants dynamically rather than by a table look up.	Yes	This option adds further flexibility to conflation of syntactic variants.
y	Attempts to disambiguate among concepts scoring equally well in matching input text by choosing concepts having the most likely semantic type in the given context.	Yes	This option supports correct interpretation of certain words, for example, "fall" used in "his pain started in April when he had a fall on his left knee" should be interpreted as "a sudden movement downward, usually resulting in injury" rather than "the season between the autumnal equinox and the winter solstice."
Y	Favors mappings with more concepts over those with fewer concepts.	No	Instead of fixed *n*-grams, we prefer to identify the longest collocationally stable word sequences, for example, a single concept "ligament tear" instead of 2 separate concepts "ligament" and "tear." In addition, longer matches also reduce ambiguity, for example, recognizing "tear" as part of "ligament tear" prevents its incorrect interpretation as "the fluid secreted by the lacrimal glands."
J	Restricts to semantic types in the comma-separated list.	Yes	To reduce the number of incorrect interpretations, we limited concept mappings to a fixed list of most relevant semantic types, which have been selected manually by a clinical expert.[b]

[a]UMLS: Unified Medical Language System.

[b]The full list of semantic types and their mappings is available from MetaMap Documentation [35].

Textbox 3. Markup of semantic types. DORA: daily or recreational activity.

1. This 22 year old was tackled in *rugby/DORA* [35] and sustained an injury.
2. She is a delightful 27 year old female who when *skiing/DORA* last year felt something pop in her knee.
3. He is normally quite active and enjoys *football/DORA*, which he is now unable to do.
4. It first started about an hour after playing *badminton/DORA*, which is something that he does.
5. He was previously very active and was involved in *sport/DORA* but has been unable to recently.
6. He is a keen *ice hockey/DORA* player.
7. Thank you for seeing this man who two years ago injured his right knee playing *basketball/DORA*.
8. She is a very athletic female, and back in 2013 had a *netball/DORA* injury.
9. It was not caused by trauma, but playing *golf/DORA* worsens it.
10. Patient is normally very fit and active playing *tennis/DORA* on a weekly basis.

Topic Modeling

To implement our topic modeling approach, we used the LDA method, which discovers latent topics in a corpus of documents based on a Bayesian statistical modeling approach [12]. This approach was chosen to support patient triage for the following reasons. By not fixing patient cohorts in advance, we wanted to avoid the need for manual annotation of data. More importantly, an unsupervised approach can identify previously unobserved patient groups beyond the boundaries of a predetermined classification scheme. Unlike cluster analysis, which can be used to support the same goal, topic modeling allows cluster overlap. This makes the problem of referring patients to multiple treatments easier to model. Interpretation of such a model is supported by (1) word distributions per topic and (2) topic distributions per document.

We used an open-source implementation of the LDA algorithm included in the Gensim library [36]. Each document was represented by a bag of words (BOW), which means that word positions and their local contexts were not taken into account. This can be partly remedied by introducing *n*-grams into the BOW representation. As described earlier, we opted to use tokens that represent markups of concepts and semantic types as an alternative to *n*-grams with added benefits of normalizing lexical and syntactic variation associated with biomedical terms. We ran experiments with different combinations of features, as described in Table 3.

Table 3. Data sets used in experiments with different types of features included.

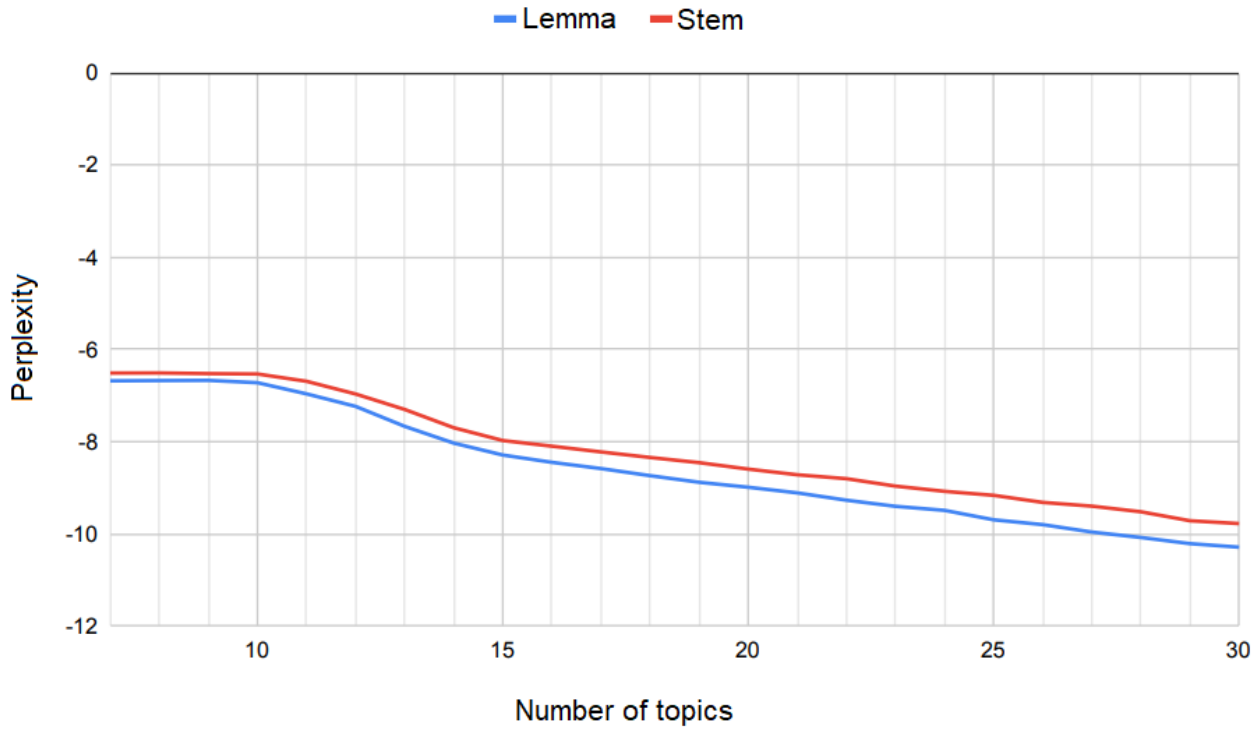
Data set	Words	Concepts	Semantic types
D1	Yes	No	No
D2	Yes	Yes	No
D3	Yes	No	Yes
D4	Yes	Yes	Yes

Hyperparameter Tuning

The performance of machine learning models depends not only on the parameters whose values the model learns during the training phase (eg, the weights for each word in a given topic) but also on the values of hyperparameters (eg, the number of topics), which are fixed before the training begins. The predictive performance of different topic modeling algorithms was found to vary substantially in practice. However, when the hyperparameters were optimized, these differences diminished significantly [37]. One of the key hyperparameters of the LDA algorithm is the number of topics. The difficulty arises when the number of relevant topics is not known a priori. An insufficient or excessive number of topics could render an LDA model too coarse or overly complex, respectively.

Perplexity, a measure of how well a probabilistic model predicts a sample, is commonly used to evaluate topic models. It is calculated as the inverse of the geometric mean per-word likelihood, with lower values indicating better models [38]. A heuristic approach based on the rate of perplexity change as a function of the number of topics has been proposed to determine an appropriate number of topics [39]. This approach would suggest selecting 11 as the total number of topics based on the values shown in Figure 3.

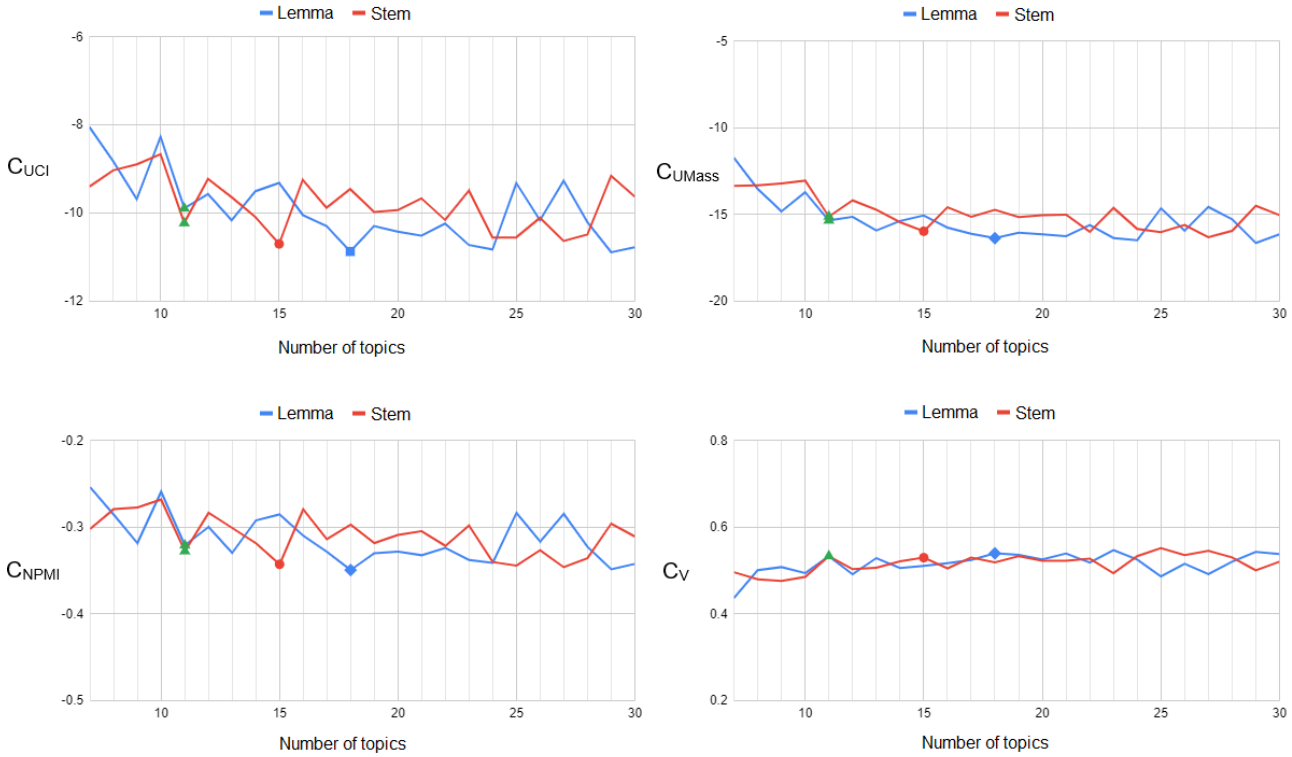
Figure 3. Perplexity as a function of the number of topics.



In general, perplexity was found not to be well correlated with the human rating of topic interpretability [40]. Alternative measures based on word coherence have been proposed to remedy this problem [41]. We used 4 measures of topic coherence, which are described in more detail in the *Results* section. As Figure 4 illustrates, the coherence of stemmed and lemmatized text achieved an optimum using 15 and 18 topics labeled by red circles and blue squares, respectively. However, at both points, topic coherence demonstrated opposite trends. However, at another local optimum labeled by green triangles,

topics modeled on stemmed and lemmatized text demonstrated not only similar trends but also almost identical coherence values. Given a small difference from the global optimum, we selected 11 as the total number of topics to be able to switch freely between stemming and lemmatization in subsequent experiments. This choice also complied with the one based on perplexity.

Figure 4. Topic coherence as a function of the number of topics.

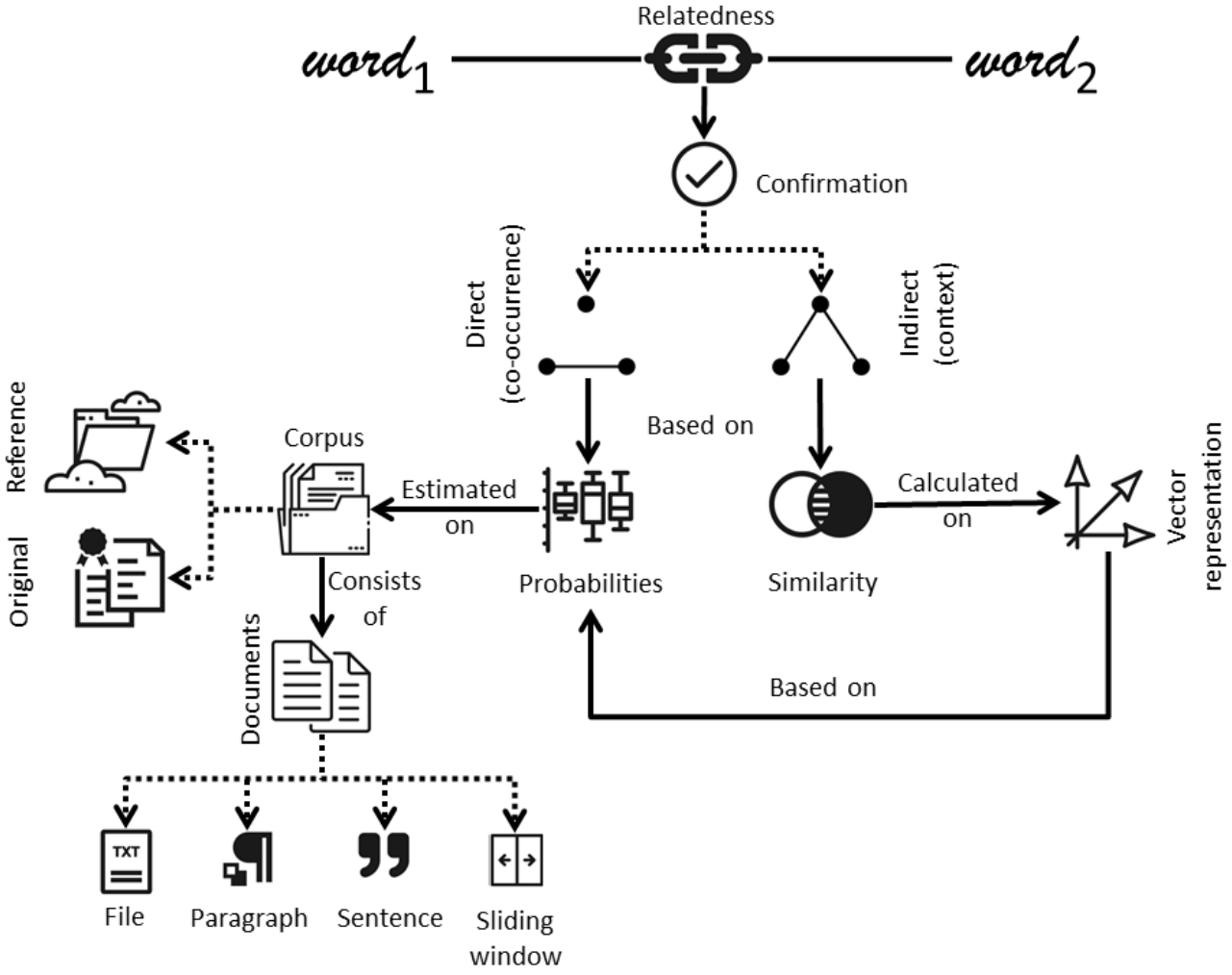


Results

Intrinsic Evaluation

Recent studies have shown that optimizing a model for perplexity may not yield human interpretable topics [40]. This limitation has prompted further research into alternative ways of estimating human interpretability. Newman et al [42] introduced the notion of topic coherence, which is based on the coherence of words that describe a topic. Different variants of this measure have been proposed [41]. In principle, overall coherence is averaged across word pairs in a topic and then across topics. Therefore, the overall topic coherence depends on the way the coherence between 2 words is measured. Figure 5 focuses on this problem. In principle, coherence refers to the degree to which 2 words are related. Two approaches to measuring relatedness can be used: one based on direct co-occurrence (or collocation) and the other based on co-occurrence with a shared set of other words.

Figure 5. Corpus-based approaches to measuring word coherence.



In the first approach, 2 words are said to be collocated if they co-occur more often than would be expected by chance. In corpus linguistics, collocation is measured by estimating relevant probabilities from a corpus of text documents, which can be either the original corpus used to learn the topic model or a reference corpus such as Wikipedia. Probabilities are estimated using Boolean documents. The number of documents in which the word (or a pair of words) occurs is divided by the total number of documents. Neither the number of occurrences within a document nor the distances between words are taken into account; hence, the name Boolean. A virtual document can be defined as a paragraph, sentence, or text window, which, by being smaller parts of the whole document, indirectly account for the distances between words.

These probabilities are used to calculate pair-wise word coherence measures such as pointwise mutual information (PMI) [43], normalized pointwise mutual information (NPMI) [44], or log-conditional probability (LCP) [45] as follows (small positive is added to avoid logarithm of zero):

PMI compares the probability of 2 words co-occurring, $P(w_i, w_j)$, against the probability that they would co-occur under the assumption of their independence, $P(wi)P(w_j)$. Higher values indicate a stronger association between the 2 words. NPMI follows the same logic, but it also imposes a fixed upper bound of 1 to indicate perfect association by normalizing PMI using the joint probability of 2 words. This makes its interpretation more intuitive while also reducing the bias toward less frequently occurring words. Both measures are symmetric, which is not a property of human word associations. By basing LCP on a simple conditional probability $P(w_i \mid w_j)$, it adds direction to measuring the association of 2 words.

Topic coherence is calculated by averaging the pair-wise word coherence across its n words:

Topic coherence measures based on PMI, NPMI, and LCP are commonly referred to as C_{UCI} (or C_{PMI}) [42], C_{NPMI} [46], and C_{UMass} [47], respectively. The problem with these measures is that they may fail to identify synonyms as related words as they do not co-occur regularly. However, we can reuse any of the

pair-wise word coherence measures to represent each word w_i as a vector whose j-th coordinate corresponds to $C(w_i, w_j)$. On the basis of the distributional hypothesis, which states that words with similar distributions have similar meanings, we can use cosine similarity between the corresponding vectors to estimate the similarity between 2 words:

Topic coherence can now be calculated by averaging the contextual similarity across its n words [46]:

In a comparative analysis, the best correlation with human topic coherence ratings was achieved with C_V [41], a topic coherence measure that uses cosine similarity on context vectors based on C_{NPMI} but differs from C_{cos} in a way in which it aggregates the similarity values. Instead of pair-wise comparison, each word is compared with the set of top-ranked words whose context vectors have been summed up.

The Gensim library [36], which was used to create topic models, was also used to calculate their coherence. It implements 4 coherence measures: C_{UCI} [42], C_{NPMI} [46], C_{UMass} [47], and C_V [41]. Table 4 reports their values obtained for topic models extracted from the data sets described in Table 3. Overall, the best results were achieved on data set D2, which was obtained by annotating the original text with concepts from the UMLS.

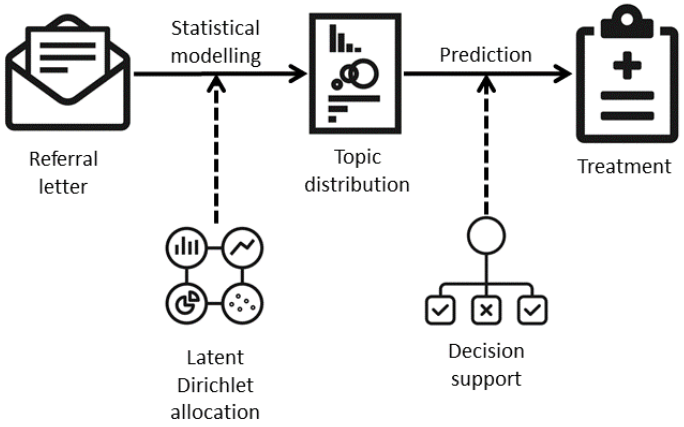
Table 4. Topic coherence.

Data set	C_{UCI}	C_{NPMI}	C_{UMass}	C_V
D1	−9.89	−0.32	−15.34	0.53
D2	−12.23	−0.41	−17.31	0.68
D3	−10.68	−0.35	−17.50	0.59
D4	−11.12	−0.37	−17.12	0.59

Extrinsic Evaluation

The extrinsic evaluation assesses the performance of a topic model in the context of a predefined task. In an envisaged scenario, topic modeling could be used to semiautomate patient triage by using topics to predict the most appropriate treatments (Figure 6). Our data set included the referral letters together with subsequently received treatments (Table 1).

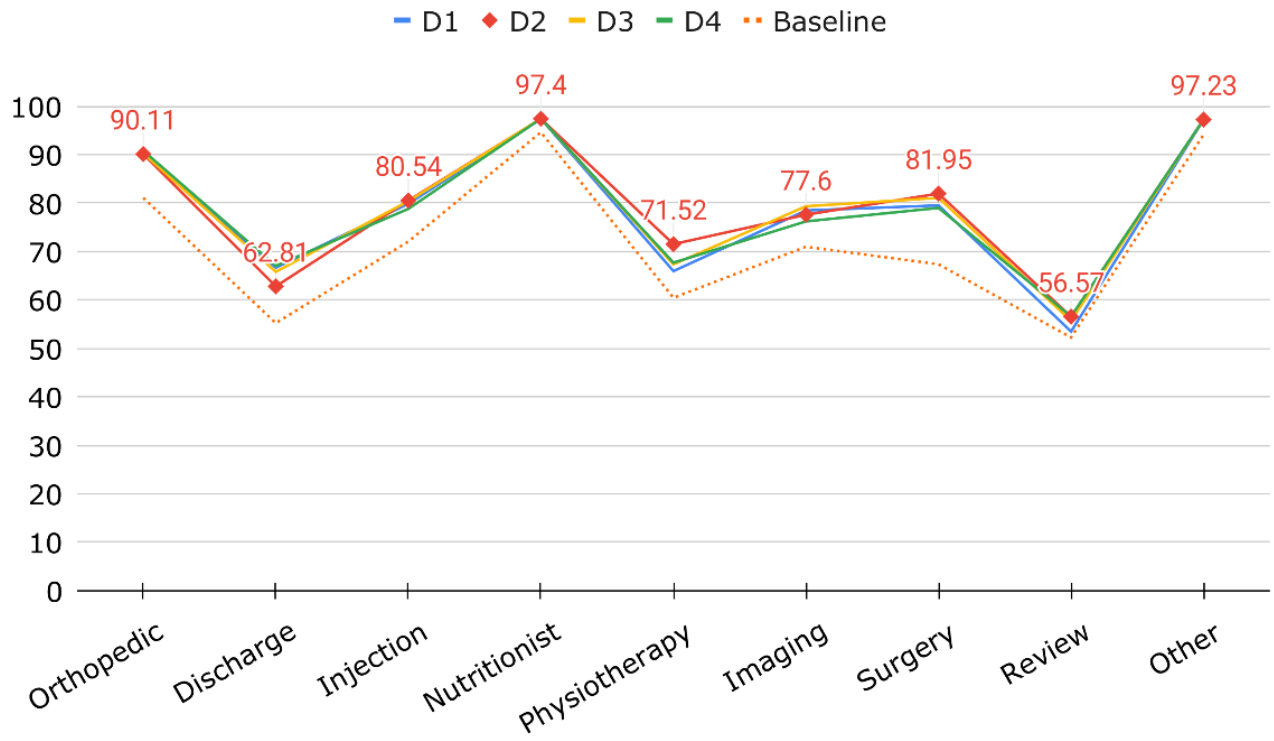
Figure 6. Supporting patient triage with topic modeling.



As a result of topic modeling, each referral letter was mapped to a topic distribution vector. Each coordinate contained a score that the letter received against the corresponding topic. Effectively, the corpus was transformed into a document-topic matrix. We trained a binary classifier for each treatment using the document-topic matrix. It takes a topic distribution vector of a referral letter as input and outputs a yes or no decision for the corresponding treatment.

We used 10-fold cross-validation to measure its prediction accuracy A=(TP+TN)/N, which was calculated using true positives (TP), true negatives (TN), and the total number (N). Cross-validation experiments were performed for each data set described in Table 3. Given a small number of features combined with few instances of some treatment outcomes, we opted for the k-nearest neighbor algorithm with k=5 in a quest to reduce overfitting. The cross-validation results are shown in Figure 7.

Figure 7. Predictive accuracy of a classifier trained on top of a topic model.

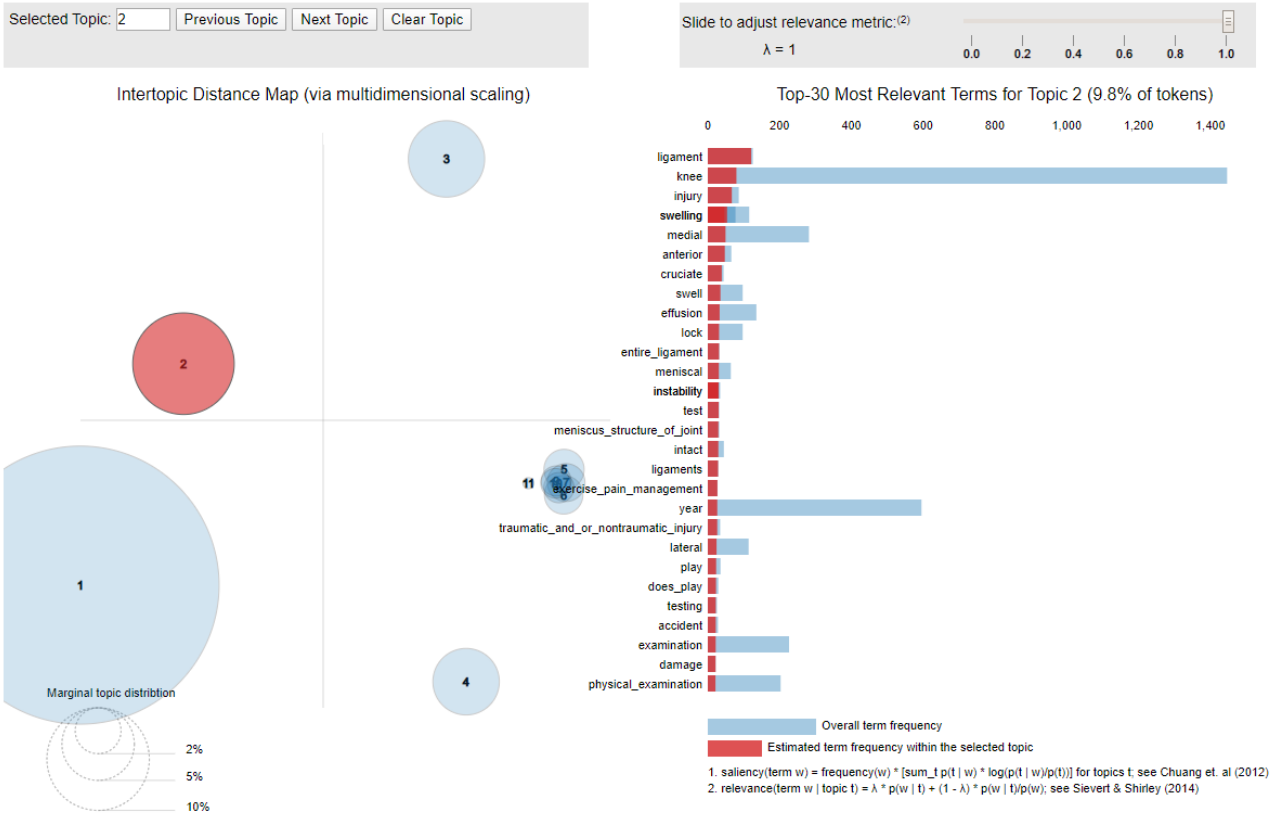


Not surprisingly, the worst results were achieved on discharge and review appointment. One would intuitively expect that these outcomes would be the least homogeneous with respect to topic distribution. In other words, any musculoskeletal patient would eventually be either discharged or reviewed, regardless of their condition. The best results were achieved for the 2 most imbalanced treatment outcomes, Nutritionist and Any other referral, with only 15 and 16 positive instances, respectively, out of a total of 576, where overfitting the majority class was most likely to have occurred. The accuracy of predicting the remaining treatment outcomes outperformed the stratified random classifier by a large margin, indicating that topic modeling could be used to support patient triage (Figure 6). On average, the best accuracy was achieved on data set D2, which augments the raw text features with domain-specific concepts. The best performance is in line with the best topic coherence recorded in the intrinsic evaluation (Table 4).

Qualitative Evaluation

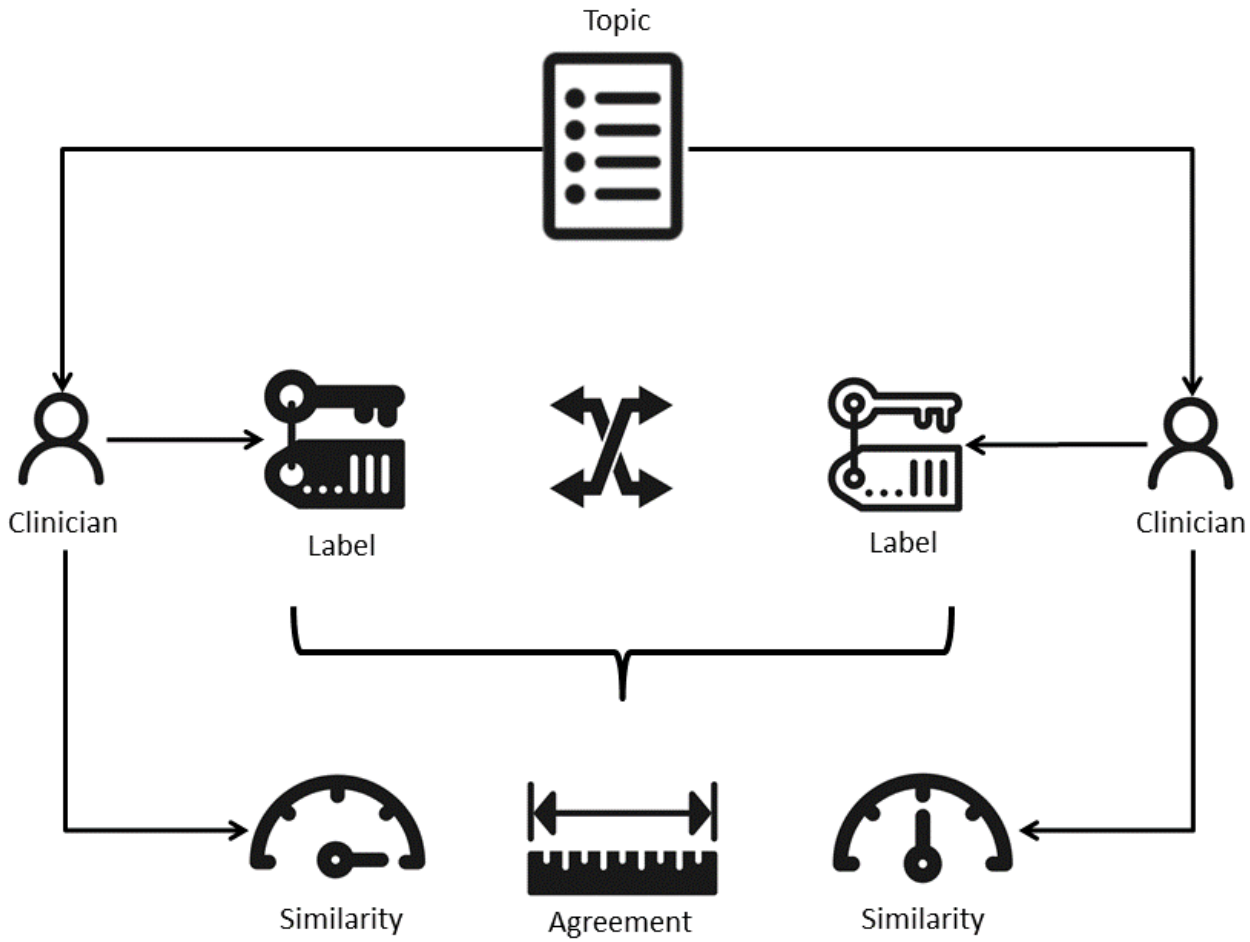
Qualitative evaluation is de facto the gold standard for measuring the interpretability of a topic model. However, involving human raters makes such an evaluation expensive to implement in practice. For that reason, we singled out a topic model with the highest coherence (Table 4) and classification accuracy (Figure 7) for further evaluation with respect to its interpretability. Its interactive web-based visualization (see Figure 8 for an example) was created using pyLDAvis, a Python library designed to help users interpret a set of latent topics [48]. Each topic was represented by a circle whose size reflects its prevalence in the training corpus. The distance between the centers of the 2 circles reflected the similarity between the corresponding topics. Clicking on a circle resulted in a histogram of the top 30 words most relevant to the corresponding topic. Here, relevance was determined based on a parameter (0 1). By default, λ was set to 1 to rank the words by their probability within a topic. When λ was set to 0, the words were reranked by their lift, which is defined as the ratio of a word's probability within a topic to its marginal probability across the corpus. The interactive interface allowed a user to adjust the value of λ between 0 and 1.

Figure 8. Interactive visualization of a topic model.



To measure the interpretability of topics, we designed experiments using a novel protocol illustrated in Figure 9. In this scenario, 2 medical doctors with specialization in physiatry were paired. Independently, each clinician was presented with an interactive visualization of the topic model (Figure 8). They completed a survey in which they were asked to describe each topic using a short free-text statement that generalizes the collective meaning of the topic's 30 most relevant words as a cohort of patients. No restrictions were imposed on the facets used in their description (eg, age, fitness, or pathology) or the choice of vocabulary. Although describing individual topics, the 2 clinicians were also asked to estimate the confidence in their final choice on a 5-point Likert scale: 0 (not confident at all), 1 (slightly confident), 2 (somewhat confident), 3 (moderately confident), and 4 (very confident).

Figure 9. Experimental protocol for measuring topic interpretability.



In the second phase, both clinicians gained access to the other one's choice of a topic's description. They were then asked to independently estimate the similarity of the 2 descriptions on a 6-point Likert scale: −3 (very dissimilar), −2 (moderately dissimilar), −1 (slightly dissimilar), 1 (slightly similar), 2 (moderately similar), and 3 (very similar). The average similarity was used to estimate the interpretability of topics under the hypothesis that high similarity implies high interpretability and vice versa. The responses to the 2 questionnaires are presented in Table 5.

Table 5. The responses to topic interpretability questionnaires.

Topic and description	Confidence	Similarity
T1		
Symptomatic degenerative conditions related to the musculoskeletal system, most commonly the knee and predominantly in females.	Moderately confident	Moderately similar
Chronic knee pain caused by an injury, causing problems for months and with a positive medical history. Related to women, medial side, and examined by x-ray. In addition to injury, chronic diseases include osteoarthritis, which can be examined by radiological diagnosis and physical examination, which reduces the range of motion and the ability to walk, and which can be treated with physical therapy and other procedures to reduce the feeling of pain.	Moderately confident	Very similar
T2		
Knee ligament injuries with a description of the type of ligament and associated symptoms, most commonly effusion.	Moderately confident	Very similar
Traumatic and nontraumatic injuries of knee ligaments, especially the medial and anterior cruciate ligaments, with swelling, effusion, and the involvement of the entire ligament leading to instability and locking of the knee. The entire ligamentous apparatus and menisci need to be tested. A history of recurrent injuries plays a role in the damage. Exercise and pain management are recommended.	Moderately confident	Very similar
T3		
Diagnosis of the pathological condition predominantly by magnetic resonance imaging together with a description of the knee injury type.	Moderately confident	Very similar
Magnetic resonance imaging used to diagnose mostly knee damage, thinning of cartilage, lateral ligaments, and hyaline and less for facets, fissures, and patellar problems.	Moderately confident	Very similar
T4		
Pathological conditions related to the hip.	Moderately confident	Moderately similar
Degenerative changes of the hip diagnosed by x-ray imaging, hip pain, decreased mobility, and reduced joint space, possibly requiring a hip replacement. Osteoarthritis diagnosed from jagged edges and anti-inflammatory processes. All these changes lead to a decreased range of motion and depression.	Somewhat confident	Very similar
T5		
Coping with sports injuries related to the musculoskeletal system.	Moderately confident	Very similar
Sports injury mostly caused by twisting. Treated with ibuprofen and bracing. Diagnosed by radiography.	Moderately confident	Very similar
T6		
Medications for painful conditions of the musculoskeletal system.	Moderately confident	Very similar
Knee injuries treated with a variety of medications.	Somewhat confident	Very similar
T7		
Musculoskeletal condition (knee) that requires an invasive procedure.	Moderately confident	Very dissimilar
Injuries that occur due to obesity and inactivity.	Slightly confident	Moderately dissimilar
T8		
Degenerative changes in the musculoskeletal system resulting in reduced activity and comorbidities.	Moderately confident	Moderately similar
Cardiovascular diseases associated with chronic lung disease, hypertension, coagulation disorder.	Somewhat confident	Slightly similar
T9		
Musculoskeletal condition (knee) more often in the female population.	Somewhat confident	Moderately similar
Most commonly, popliteal cyst, a predisposition in occupations that require prolonged standing, can lead to knee deformities. Excision is a recommended treatment.	Slightly confident	Very similar

Topic and description	Confidence	Similarity
T10		
Pain in the lumbosacral spine.	Somewhat confident	Very similar
Changes in the lumbar spine and pelvis due to osteoarthritis and infection. Accompanied by hot, burning back pain and progression.	Slightly confident	Very similar
T11		
Patients with amputation of the lower extremities.	Moderately confident	Very dissimilar
Poor mobility due to asymmetries.	Slightly confident	Very dissimilar

The average confidence was found to be 3.00 and 2.00 between the two annotators. The average similarity was found to be 2.00 for both annotators. One participant was consistently more confident than the other, but they were mostly not more than one Likert point apart. The biggest discrepancy between the 2 Likert points was found for topics T8 and T11. When cross-referenced against the topic similarity scores, most dissimilar descriptions were observed. Overall, the participants' perception of topic similarity was consistent, with one Likert point difference throughout.

To generalize these findings, we calculated the interannotator agreement for both confidence and similarity (Table 6). For this purpose, we used Cohen kappa coefficient with linear weighting [49-52]. The agreement on confidence was low. However, a closer look at the distribution of confidence scores between the 2 participants revealed that one participant was consistently more confident than the other. Therefore, the low agreement on confidence in interpreting the topics was more likely to be associated with the participants' own characteristics than the topics themselves. Indeed, the participant with higher confidence provided more generic descriptions, whereas the other paid more attention to detail, which may have lowered their confidence in believing that they addressed the task effectively. Nonetheless, in the vast majority of cases (9 out of 11 topics), the high similarity scores indicate that both generic and detailed descriptions effectively referred to the same cohort, that is, a group of patients who share common characteristics or experiences such as medical history, demographics, and possible treatments. Therefore, based on the hypothesis that high similarity implies high interpretability and vice versa, we conclude that the given topic model was highly interpretable.

Table 6. Interannotator agreement on topic description.

Characteristics	Confidence	Similarity
Observed kappa	0.1391	0.7343
Standard error	0.0925	0.1163
Confidence interval	0.0000-0.3204	0.5063-0.9623
Maximum possible	0.1391	0.7343
Proportion of maximum possible	1	1

Discussion

Principal Findings

This study explored the feasibility of using NLP and machine learning to automate triage of patients with musculoskeletal conditions by analyzing information from referral letters. Specifically, we determined that LDA can automatically assort referral letters into topics that are clinically relevant. In other words, latent topics provide information that is considered relevant when prescribing treatments.

First, our experiments confirmed that latent topics could be used to automatically predict an appropriate treatment. A supervised classifier based on latent topics as its sole feature consistently outperformed the baseline method. Further improvements in the performance of such classifiers stand to be gained by incorporating other types of features that can be obtained from the patients' electronic health records, for example, demographics, body mass index, and imaging reports. However, this was beyond the scope of this study, which was concerned only with establishing the clinical relevance of automatically extracted latent topics. On their own, these topics proved to be sufficiently discriminative features for treatment recommendations based on machine learning.

Second, our experiments confirmed that latent topics could be interpreted by clinicians as cohorts of patients who share common characteristics or experiences such as medical history, demographics, and possible treatments. Specifically, the words associated with each topic by the LDA algorithm proved to be sufficiently descriptive to enable clinical specialists to interpret the topic's underlying semantics.

The first set of experiments established the clinical relevance of latent topics from a machine perspective: a treatment can be recommended automatically for an individual patient. The second set of experiments established the clinical relevance of

latent topics from a human perspective: a treatment can be recommended by a clinician for an automatically identified cohort of patients. Both treatment recommendation scenarios support the hypothesis that topic modeling can support patient triage. Automating this process can be used to address areas where bottlenecks exist. Efficient referral to appropriate services such as analgesia or diagnostics not only improves patient experience and health outcomes but also reduces queuing arising from nonurgent demand, thus minimizing the delays for those with urgent care needs.

Conclusions

Our approach used information contained in referral letters to underpin the referral decision-making process. Successful automation of this process has the potential to streamline care pathways and ensure that patients receive timely and optimal care. In clinical applications such as patient triage, interpretability is the key to build trust for all stakeholders, clinicians, and patients alike. Our approach to qualitative evaluation sets a precedent in measuring the interpretability of automated outputs, which is emerging as the next big challenge for clinical NLP. The unsupervised aspect of the proposed approach avoids the need for data annotation and, therefore, can be readily deployed to tackle other bottlenecks along the musculoskeletal pathway. For example, imaging and pathology reports can be processed in the same way to automatically redirect patients to the most appropriate services.

Acknowledgments

This study was partly funded by Health and Care Research Wales as part of a project on improving access to care and treatment for patients with hip and knee pain at the primary and secondary care interface (ref: RfPPB 1114). The authors are grateful to Dr Tatjana Knezevic and Dr Dejan Nikolic from the University Hospital in Belgrade, Serbia, for providing feedback on the clinical interpretability of the topic model.

Authors' Contributions

KB and IS designed the study. IS designed and implemented the system. KB coordinated data collection and qualitative evaluation. IS drafted the manuscript. Both authors contributed to the sections related to their involvement in the study. Both authors reviewed and approved the manuscript for publication.

Conflicts of Interest

None declared.

References

1. Musculoskeletal Conditions. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/conditions-and-diseases/musculoskeletal-conditions [accessed 2020-01-01]
2. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2163-2196 [FREE Full text] [doi: 10.1016/S0140-6736(12)61729-2] [Medline: 23245607]
3. Johnson SA, Kalairajah Y, Moonot P, Steele N, Field RE. Fast-track assessment clinic: selection of patients for a one-stop hip assessment clinic. Ann R Coll Surg Engl 2008 Apr;90(3):208-212 [FREE Full text] [doi: 10.1308/003588408X242024] [Medline: 18430334]
4. Inglis T, Armour P, Inglis G, Hooper G. Rationing of hip and knee referrals in the public hospital: the true unmet need. N Z Med J 2017 Mar 24;130(1452):39-48. [Medline: 28337039]
5. Spasić I, Zhao B, Jones C, Button K. KneeTex: an ontology-driven system for information extraction from MRI reports. J Biomed Semantics 2015;6:34 [FREE Full text] [doi: 10.1186/s13326-015-0033-1] [Medline: 26347806]
6. Hassanpour S, Langlotz C, Amrhein T, Befera N, Lungren M. Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield. AJR Am J Roentgenol 2017 Apr;208(4):750-753. [doi: 10.2214/AJR.16.16128] [Medline: 28140627]
7. Spasić I, Owen D, Smith A, Button K. KLOSURE: closing in on open-ended patient questionnaires with text mining. J Biomed Semantics 2019 Nov 12;10(Suppl 1):24 [FREE Full text] [doi: 10.1186/s13326-019-0215-3] [Medline: 31711536]
8. Spasić I, Uzuner O, Zhou L. Emerging clinical applications of text analytics. Int J Med Inform 2020 Feb;134:103974. [doi: 10.1016/j.ijmedinf.2019.103974] [Medline: 31630961]
9. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014 Sep;83(9):605-623 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.06.009] [Medline: 25008281]
10. Spasic I, Nenadic G. Clinical text data in machine learning: systematic reviews. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
11. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):e1312 [FREE Full text] [doi: 10.1002/widm.1312] [Medline: 32089788]

12. Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. J Mach Learn Res 2003;3:993-1022 [FREE Full text]
13. Huang Z, Lu X, Duan H. Latent treatment pattern discovery for clinical processes. J Med Syst 2013 Apr;37(2):9915. [doi: 10.1007/s10916-012-9915-2] [Medline: 23389419]
14. Arnold CW, Oh A, Chen S, Speier W. Evaluating topic model interpretability from a primary care physician perspective. Comput Methods Programs Biomed 2016 Feb;124:67-75 [FREE Full text] [doi: 10.1016/j.cmpb.2015.10.014] [Medline: 26614020]
15. Wang L, Wang Y, Shen F, Rastegar-Mojarad M, Liu H. Discovering associations between problem list and practice setting. BMC Med Inform Decis Mak 2019 Apr 4;19(Suppl 3):69 [FREE Full text] [doi: 10.1186/s12911-019-0779-y] [Medline: 30943957]
16. Rumshisky A, Ghassemi M, Naumann T, Szolovits P, Castro VM, McCoy TH, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Transl Psychiatry 2016 Oct 18;6(10):e921 [FREE Full text] [doi: 10.1038/tp.2015.182] [Medline: 27754482]
17. Fong A, Ratwani R. An evaluation of patient safety event report categories using unsupervised topic modeling. Methods Inf Med 2015;54(4):338-345. [doi: 10.3414/ME15-01-0010] [Medline: 25833655]
18. Bisgin H, Liu Z, Fang H, Xu X, Tong W. Mining FDA drug labels using an unsupervised learning technique--topic modeling. BMC Bioinformatics 2011 Oct 18;12(Suppl 10):S11 [FREE Full text] [doi: 10.1186/1471-2105-12-S10-S11] [Medline: 22166012]
19. Sullivan R, Sarker A, O'Connor K, Goodin A, Karlsrud M, Gonzalez G. Finding Potentially Unsafe Nutritional Supplements From User Reviews With Topic Modeling. In: Pacific Symposium on Biocomputing. 2016 Presented at: PSB'16; December 11-17, 2016; Hawaii, USA p. 528-539 URL: https://www.worldscientific.com/doi/abs/10.1142/9789814749411_0048 [doi: 10.1142/9789814749411_0048]
20. Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, Kho A, et al. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform 2015 Jun;55:82-93 [FREE Full text] [doi: 10.1016/j.jbi.2015.03.011] [Medline: 25841328]
21. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology 2018 May;287(2):570-580. [doi: 10.1148/radiol.2018171093] [Medline: 29381109]
22. Barroilhet SA, Pellegrini AM, McCoy TH, Perlis RH. Characterizing DSM-5 and ICD-11 personality disorder features in psychiatric inpatients at scale using electronic health records. Psychol Med 2020 Oct;50(13):2221-2229. [doi: 10.1017/S0033291719002320] [Medline: 31544723]
23. Button K, Spasić I, Playle R, Owen D, Lau M, Hannaway L, et al. Using routine referral data for patients with knee and hip pain to improve access to specialist care. BMC Musculoskelet Disord 2020 Feb 3;21(1):66 [FREE Full text] [doi: 10.1186/s12891-020-3087-x] [Medline: 32013997]
24. Spasic I, Krzeminski D, Corcoran P, Balinsky A. Cohort selection for clinical trials from longitudinal patient records: text mining approach. JMIR Med Inform 2019 Oct 31;7(4):e15980 [FREE Full text] [doi: 10.2196/15980] [Medline: 31674914]
25. Griffis D, Shivade C, Fosler-Lussier E, Lai A. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. AMIA Jt Summits Transl Sci Proc 2016;2016:88-97 [FREE Full text] [Medline: 27570656]
26. Pakhomov S, Pedersen T, Chute C. Abbreviation and acronym disambiguation in clinical discourse. AMIA Annu Symp Proc 2005:589-593 [FREE Full text] [Medline: 16779108]
27. Spasić I, Greenwood M, Preece A, Francis N, Elwyn G. FlexiTerm: a flexible term recognition method. J Biomed Semantics 2013 Oct 10;4(1):27 [FREE Full text] [doi: 10.1186/2041-1480-4-27] [Medline: 24112363]
28. Spasic I. Acronyms as an integral part of multi-word term recognition – a token of appreciation. IEEE Access 2018;6:8351-8363 [FREE Full text] [doi: 10.1109/access.2018.2807122]
29. Documentation. NLTK Project. URL: https://www.nltk.org/api/nltk.chunk.html [accessed 2020-01-01]
30. Named Entity Recognition. Explosion AI. URL: https://spacy.io/api/annotation#named-entities [accessed 2020-01-01]
31. Button K, van Deursen RW, Soldatova L, Spasić I. TRAK ontology: defining standard care for the rehabilitation of knee conditions. J Biomed Inform 2013 Aug;46(4):615-625 [FREE Full text] [doi: 10.1016/j.jbi.2013.04.009] [Medline: 23665300]
32. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
33. Wang X, McCallum A, Wei X. Topical N-grams: Phrase and Topic Discovery, With an Application to Information Retrieval. In: Seventh IEEE International Conference on Data Mining. 2007 Presented at: CDM'07; September 3-9, 2007; Omaha, USA. [doi: 10.1109/icdm.2007.86]
34. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
35. Semantic Type Mappings. MetaMap Documentation. 2020. URL: https://metamap.nlm.nih.gov/Docs/SemanticTypes_2018AB.txt [accessed 2020-01-01]
36. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: LREC Workshop on New Challenges for NLP Frameworks. 2010 Presented at: NLP'10; June 2-8, 2010; Valletta, Malta URL: https://is.muni.cz/publication/884893/en/Software-Framework-for-Topic-Modelling-with-Large-Corpora/Rehurek-Sojka

37. Asuncion A, Welling M, Smyth P, Teh Y. On Smoothing and Inference for Topic Models. In: 25th Conference on Uncertainty in Artificial Intelligence. 2009 Presented at: UAI'09; July 22-29, 2009; Montreal, Canada URL: https://dl.acm.org/doi/10.5555/1795114.1795118
38. Wallach H, Murray I, Salakhutdinov R, Mimno D. Evaluation Methods for Topic Models. In: 26th Annual International Conference on Machine Learning. 2009 Presented at: CML'09; September 1-9, 2009; Montreal, Canada URL: https://dl.acm.org/doi/10.1145/1553374.1553515 [doi: 10.1145/1553374.1553515]
39. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics 2015;16(Suppl 13):S8 [FREE Full text] [doi: 10.1186/1471-2105-16-S13-S8] [Medline: 26424364]
40. Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei D. Reading Tea Leaves: How Humans Interpret Topic Models. In: 22nd International Conference on Neural Information Processing Systems. 2009 Dec Presented at: NIPS'09; December 22-24, 2009; Vancouver, Canada URL: https://dl.acm.org/doi/10.5555/2984093.2984126
41. Roder M, Both A, Hinneburg A. Exploring the Space of Topic Coherence Measures. In: 8th ACM International Conference on Web Search and Data Mining. 2015 Presented at: WSDM'15; June 1-8, 2015; New York, USA URL: https://dl.acm.org/doi/10.1145/2684822.2685324
42. Newman D, Lau JH, Grieser K, Baldwin T. Automatic Evaluation of Topic Coherence. In: Conference of the North American Chapter of the Association for Computational Linguistics. 2010 Presented at: ACL'10; June 12-17, 2010; Los Angeles, USA URL: https://dl.acm.org/doi/10.5555/1857999.1858011
43. Church K, Hanks P. Word Association Norms, Mutual Information, and Lexicography. In: 27th Annual Meeting of the Association for Computational Linguistics. 1989 Presented at: ACL'89; June 1-7, 1989; Vancouver, Canada URL: https://www.aclweb.org/anthology/P89-1010/ [doi: 10.3115/981623.981633]
44. Bouma G. Normalized (Pointwise) Mutual Information in Collocation Extraction. In: Conference of the German Society for Computational Linguistics and Language Technology. 2009 Presented at: CLLT'09; July 23-27, 2009; Potsdam, Germany URL: https://www.semanticscholar.org/paper/Normalized-(pointwise)-mutual-information-in-Bouma/15218d9c029cbb903ae7c729b2c644c24994c201
45. Michelbacher L, Evert S, Schütze H. Asymmetric Association Measures. In: International Conference on Recent Advances in Natural Language Processing. 2007 Presented at: NLP'07; March 3-9, 2007; Borovets, Bulgaria URL: http://www.stefan-evert.de/PUB/MichelbacherEtc2007.pdf
46. Aletras N, Stevenson M. Evaluating Topic Coherence Using Distributional Semantics. In: 10th International Conference on Computational Semantics. 2013 Presented at: CCS'13; November 3-7, 2013; Potsdam, Germany URL: https://www.aclweb.org/anthology/W13-0102.pdf
47. Mimno D, Wallach H, Talley E, Leenders M, McCallum A. Optimizing Semantic Coherence in Topic Models. In: Conference on Empirical Methods in Natural Language Processing. 2011 Presented at: NLP'11; October 22-26, 2011; Edinburgh, UK URL: https://www.aclweb.org/anthology/D11-1024/
48. Sievert C, Shirley K. A Method for Visualizing and Interpreting Topics. In: Workshop on Interactive Language Learning, Visualization and Interfaces. 2014 Presented at: ILLVI'14; December 4-6, 2014; Baltimore, USA URL: https://www.aclweb.org/anthology/W14-3110/ [doi: 10.3115/v1/w14-3110]
49. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure 2016 Jul 2;20(1):37-46 [FREE Full text] [doi: 10.1177/001316446002000104]
50. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968 Oct;70(4):213-220. [doi: 10.1037/h0026256] [Medline: 19673146]
51. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969;72(5):323-327. [doi: 10.1037/h0028106]
52. Fleiss J, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Measure 1973;33:613-619 [FREE Full text]

Abbreviations

BOW: bag of words
CUI: concept unique identifier
LCP: log-conditional probability
LDA: latent Dirichlet allocation
NLP: natural language processing
NPMI: normalized pointwise mutual information
PMI: pointwise mutual information
TN: true negatives
TP: true positives
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 15.06.20; peer-reviewed by W Raghupathi, M Vugts, J Li, F Palmieri; comments to author 28.07.20; revised version received 17.09.20; accepted 05.10.20; published 06.11.20.

Please cite as:
Spasic I, Button K
Patient Triage by Topic Modeling of Referral Letters: Feasibility Study
JMIR Med Inform 2020;8(11):e21252
URL: https://medinform.jmir.org/2020/11/e21252
doi:10.2196/21252
PMID:33155985

Original Paper

Multidimensional Machine Learning Personalized Prognostic Model in an Early Invasive Breast Cancer Population-Based Cohort in China: Algorithm Validation Study

Xiaorong Zhong[1*], MD; Ting Luo[1*], MD; Ling Deng[2], MD, PhD; Pei Liu[3], MSc; Kejia Hu[4], MD; Donghao Lu[4], MD, PhD; Dan Zheng[2], MD; Chuanxu Luo[2], MD; Yuxin Xie[1], MD; Jiayuan Li[5], PhD; Ping He[1], MD; Tianjie Pu[6], MD; Feng Ye[6], PhD; Hong Bu[6], MD, PhD; Bo Fu[3], PhD; Hong Zheng[2], MD, PhD

[1]Department of Head, Neck and Mammary Gland Oncology, Cancer Center, West China Hospital, Sichuan University, Chengdu, China

[2]Laboratory of Molecular Diagnosis of Cancer, Clinical Research Center for Breast, West China Hospital, Sichuan University, Chengdu, China

[3]Big Data Research Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

[4]Department of Medical Epidemiology & Biostatistics, Karolinska Institutet, Stockholm, Sweden

[5]Department of Epidemiology and Biostatistics, West China School of Public Health, Sichuan University, Chengdu, China

[6]Laboratory of Pathology, West China Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Hong Zheng, MD, PhD
Laboratory of Molecular Diagnosis of Cancer
Clinical Research Center for Breast, West China Hospital
Sichuan University
37 Guoxuexiang, Wuhou District
Chengdu
China
Phone: 86 2885422685
Email: hzheng@scu.edu.cn

Abstract

Background: Current online prognostic prediction models for breast cancer, such as Adjuvant! Online and PREDICT, are based on specific populations. They have been well validated and widely used in the United States and Western Europe; however, several validation attempts in non-European countries have revealed suboptimal predictions.

Objective: We aimed to develop an advanced breast cancer prognosis model for disease progression, cancer-specific mortality, and all-cause mortality by integrating tumor, demographic, and treatment characteristics from a large breast cancer cohort in China.

Methods: This study was approved by the Clinical Test and Biomedical Ethics Committee of West China Hospital, Sichuan University on May 17, 2012. Data collection for this project was started in May 2017 and ended in March 2019. Data on 5293 women diagnosed with stage I to III invasive breast cancer between 2000 and 2013 were collected. Disease progression, cancer-specific mortality, all-cause mortality, and the likelihood of disease progression or death within a 5-year period were predicted. Extreme gradient boosting was used to develop the prediction model. Model performance was assessed by calculating the area under the receiver operating characteristic curve (AUROC), and the model was calibrated and compared with PREDICT.

Results: The training, test, and validation sets comprised 3276 (499 progressions, 202 breast cancer-specific deaths, and 261 all-cause deaths within 5-year follow-up), 1405 (211 progressions, 94 breast cancer-specific deaths, and 129 all-cause deaths), and 612 (109 progressions, 33 breast cancer-specific deaths, and 37 all-cause deaths) women, respectively. The AUROC values for disease progression, cancer-specific mortality, and all-cause mortality were 0.76, 0.88, and 0.82 for training set; 0.79, 0.80, and 0.83 for the test set; and 0.79, 0.84, and 0.88 for the validation set, respectively. Calibration analysis demonstrated good agreement between predicted and observed events within 5 years. Comparable AUROC and calibration results were confirmed in different age, residence status, and receptor status subgroups. Compared with PREDICT, our model showed similar AUROC and improved calibration values.

Conclusions: Our prognostic model exhibits high discrimination and good calibration. It may facilitate prognosis prediction and clinical decision making for patients with breast cancer in China.

(JMIR Med Inform 2020;8(11):e19069) doi:10.2196/19069

KEYWORDS

breast cancer; prognosis; machine learning; prediction model

Introduction

Breast cancer is a heterogeneous disease with different prognoses. Traditional prognostic factors include tumor size, number of positive lymph nodes, tumor grade, and molecular biomarkers such as estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and Ki67 [1].

Several prognostic prediction models have recently been developed to assist clinical decision making in breast cancer treatment [2]. These models focused on clinical and pathological factors, as well as gene expression (Oncotype, MammaPrint, BCI, and EndoPredict) [3-8]. Among the prediction models based on clinical and pathological factors, Adjuvant! Online and PREDICT are commonly used [3,4]; however, both of these models are largely based on Caucasian populations, and several validation attempts have revealed suboptimal predictions [2,9-13]. Recently, Wu et al [14] developed a race-specific breast cancer recurrence and survival model but with very few Asians. Therefore, the current models, which are based on specific populations, are inadequate for clinical practice and cannot explain the sizable variability in patient prognosis.

In this study, we aimed to develop a comprehensive prediction model for the prognosis of early invasive breast cancer using machine-learning methods. Our study was based on a large cohort of Chinese patients with breast cancer from West China Hospital, Sichuan University.

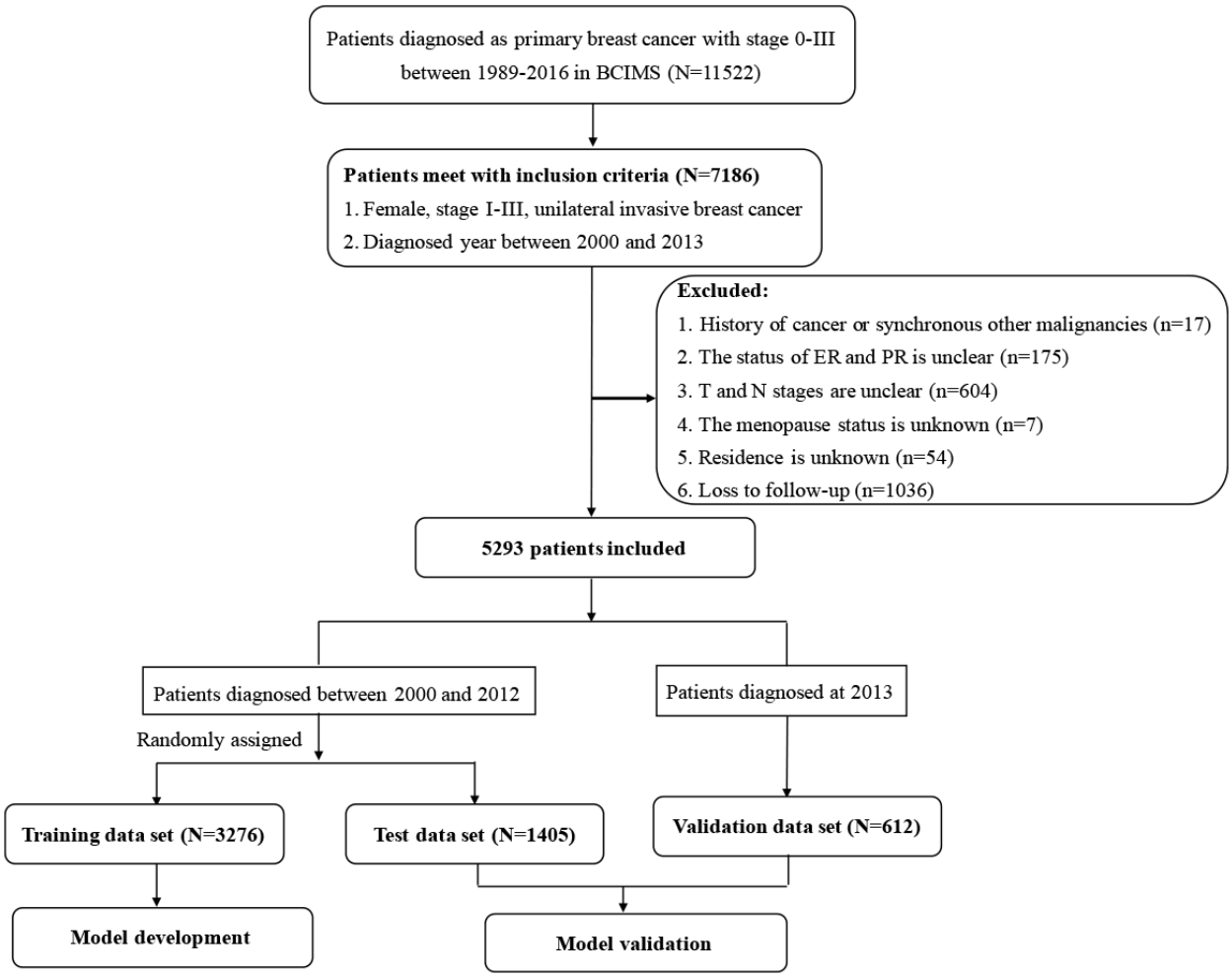
Methods

Patient Population

Patients records were derived from the Breast Cancer Information Management System (BCIMS) at the West China Hospital of Sichuan University [15]; the cases derived from the BCIMS are representative of breast cancer cases in Southwest China [16]. The BCIMS contains over 16,000 breast cancer patient cases dating back to 1989 and prospectively records patient clinical and pathological characteristics, medical history, diagnosis, laboratory results, and treatments [16].

This cohort study included women diagnosed with unilateral stage I to III invasive primary breast cancer who had undergone primary breast cancer treatment between 2000 and 2013. Patients with a history of cancer, with other synchronous malignancies, lacking important information (ER, PR, T stage, N stage, menopause status, and residence), or lost to follow-up were excluded from the study. A flow chart of the study design (with inclusions and exclusions) is shown in Figure 1. In total, 5293 patients were included. Patients diagnosed between 2000 and 2012 were randomly divided into a training set (n=3276) for model development and a test set (n=1405), for model validation, whereas those diagnosed in 2013 were used as a data set (n=612) for model validation in a separate population.

Figure 1. Flowchart of the study design and patient selection. BCIMS: Breast Cancer Information Management System; ER: estrogen receptor; PR: progesterone receptor.



Outcomes

The patients were prospectively followed using BCIMS records. Follow-up investigations, namely physical examinations, blood tests, breast ultrasounds, computed tomography, and magnetic resonance scans of the chest and abdomen were performed every 3 months for the first 2 years after surgery, then every 6 months from 3 to 5 years after diagnosis, and every year thereafter. Follow-up was conducted via interviews during outpatient visits, or by telephone or postal contact by research assistants.

The endpoints were disease progression (recurrence, metastasis, second primary tumor, and death), cancer-specific mortality (death due to breast cancer), and all-cause mortality. The likelihood of disease progression or death within a 5-year period was predicted. Patients who were alive and showed no evidence of recurrence during the 5 years of follow-up were censored at the fifth year for model development. Invasive disease-free survival was defined as the time from the date of diagnosis to the date of first documented recurrence, the date of death, or 5 years after diagnosis, whichever was earlier. Breast cancer-specific survival was defined as the time from the date of diagnosis to the date of death due to breast cancer or 5 years after diagnosis, whichever was earlier.

Statistical Analysis

Statistical analyses and modeling were performed using Python (version 3.6.2, Python Software Foundation), XGBoost (version 0.82), and STATA (version 14; Stata Corp LLC) software packages. A chi-square test was used to test the difference in the categorical variables between the training and test data sets. extreme gradient boosting (XGBoost) was used to develop the prognostic prediction model. The process of model development had 2 parts: stratified feature selection and survival modeling. Stratified feature selection has previously been described [17]. Briefly, after setting standards and cleaning the data, 39 original features were obtained to construct prognosis models (Multimedia Appendix 1). Kolmogorov–Smirnov and chi-square tests were preliminarily used to determine whether each feature, as a single factor, was significantly associated with one or more outcomes. This step selected 26 features with notable effects on outcomes. Subsequently, the XGBoost classifier was run to obtain the average importance score of each feature by performing 10-fold cross-validation 5 times with hyperparameter optimization. In this step, the weight method was applied to compute the importance score, which was the number of times a feature was used to split the data across all trees. Subsequently, subsets of features were used to find the threshold score by applying backward selection step-by-step to determine whether

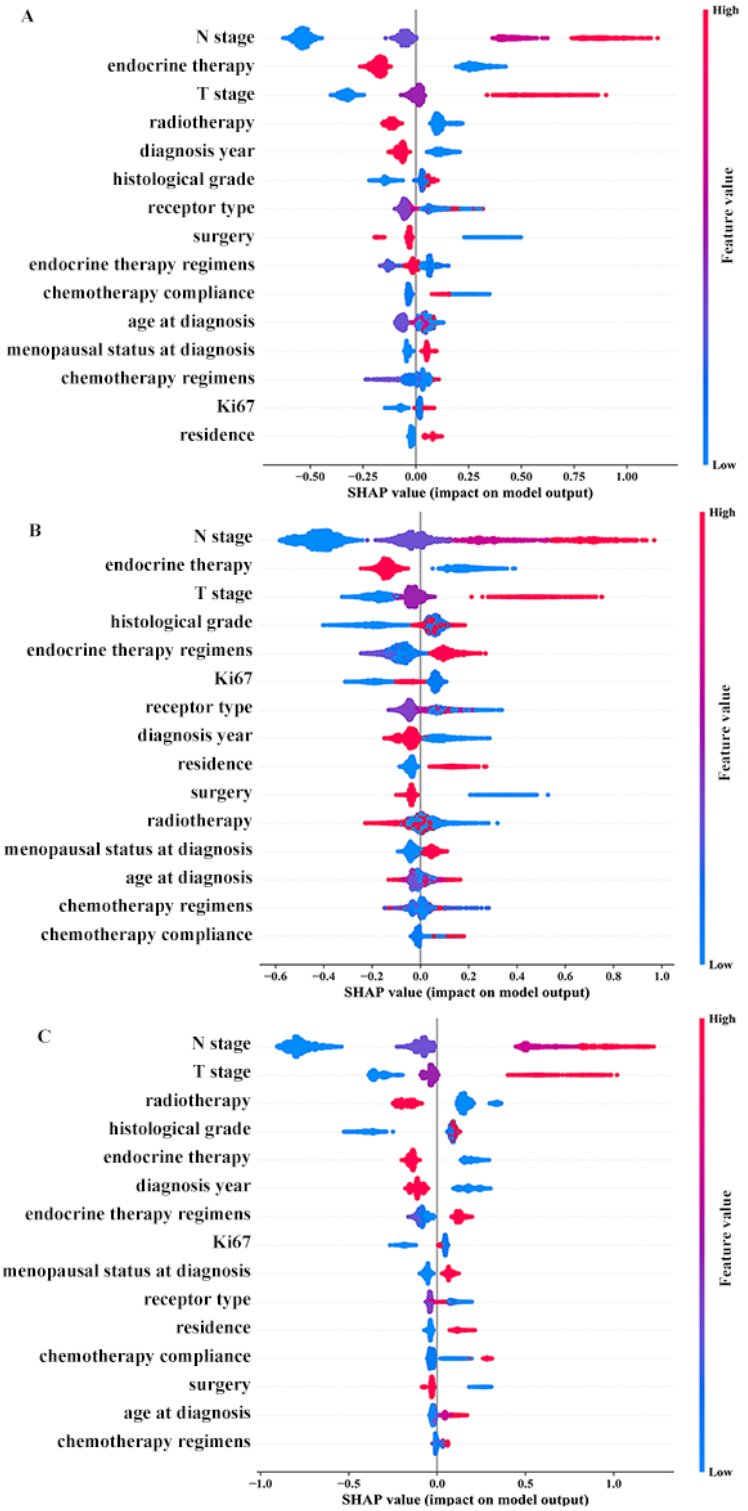
a feature score was important. The threshold score was 0.020 for disease progression, 0.015 for cancer-specific mortality, and 0.020 for all-cause mortality. Features with scores lower than the threshold score or with high similarity to other features were excluded. However, menopausal status at diagnosis, which was related to treatment and prognosis in clinical practice, was included, although it scored slightly lower than the threshold. In total, 15 variables were selected for model development (Multimedia Appendix 2). The XGBoost decision tree algorithm was used to estimate the hazard ratio, and hyperparameters were obtained using Bayesian optimization and cross-validation [18]. The likelihood of disease progression or death within a 5-year period was estimated using the equation $\hat{y}(t, X) = 1 - [S_0(t)]^{hr(X)}$, where, t denotes the observed period, X denotes the selected variables, $S_0(t)$ denotes a population-level baseline survival function, and $hr()$ denotes the hazard ratio outputted by the model, respectively. Taking into account the calibration results of the decision tree model, the estimated likelihood was further calibrated using isotonic regression (scikit-learn package, version 0.20.3) [19].

To visualize the contributions of the features in the machine learning model, Shapley additive explanations (SHAP) (shap package, version 0.28.5) and partial dependence plots (PDPbox package, version 0.2.0) were used to evaluate how each feature affected the model prediction. The SHA*P* value represents the effect of changes in a feature on the model output. By pooling the features of all samples in the training data set, the SHA*P* value plot provides an overview of the features that are most important for the model, and features on the plot are sorted by the sum of SHA*P* value magnitudes over all samples [20]. The partial dependence plot takes a row of the data set and repeatedly changes the value for the feature. This is done multiple times with different rows and then aggregated to determine how the feature affects the outcome over a wide range. A partial dependence plot is then created to show how the outcome changes with different values [21].

We compared machine learning models incorporating different variables. We also compared the machine learning model with Cox proportional hazards regression models using the same variables. For this purpose, 4 models were developed: (1) a full model with XGBoost incorporating demographic, tumor, and treatment variables (Figure 2 and Multimedia Appendix 3-5); (2) model A with XGBoost incorporating demographic and tumor variables (Multimedia Appendix 6); (3) model B with XGBoost incorporating variables similar to those in other published models (Multimedia Appendix 7, Multimedia Appendix 8) [3,4]; (4) model C with Cox incorporating the same variables as those in the full model (Multimedia Appendix 9).

Model discrimination was evaluated by generating receiver operating characteristic curves and estimating the area under the receiver operating characteristic curves (AUROC) for the models. The DeLong test was used to compare the AUROC values between the models. The predicted and observed 5-year events were compared for each model, and a test of proportion was used for determining the equality between predicted and observed events [14]. A calibration plot was generated using each decile of the predicted value. To explain the different states of breast cancer patients, the model performance was assessed in subgroups of different demographic and tumor characteristics. Our model was also compared with the PREDICT model [4] using test and validation data sets (Multimedia Appendix 7). All statistical tests were 2-sided unless stated otherwise, and a P value<.05 was considered statistically significant.

Figure 2. The importance of features for (A) disease progression, (B) breast cancer mortality, and (C) all-cause mortality. SHAP: Shapley additive explanation.



Ethics

This study was approved by the Clinical Test and Biomedical Ethics Committee of West China Hospital, Sichuan University (reference number 2012-130) on May 17, 2012. Written informed consent was provided by each patient. Data collection started in May 2017 and ended in March 2019. A total of 11,522 patients were recruited, and 5293 patients were included in the analysis for model development and validation.

Results

Study Population Characteristics

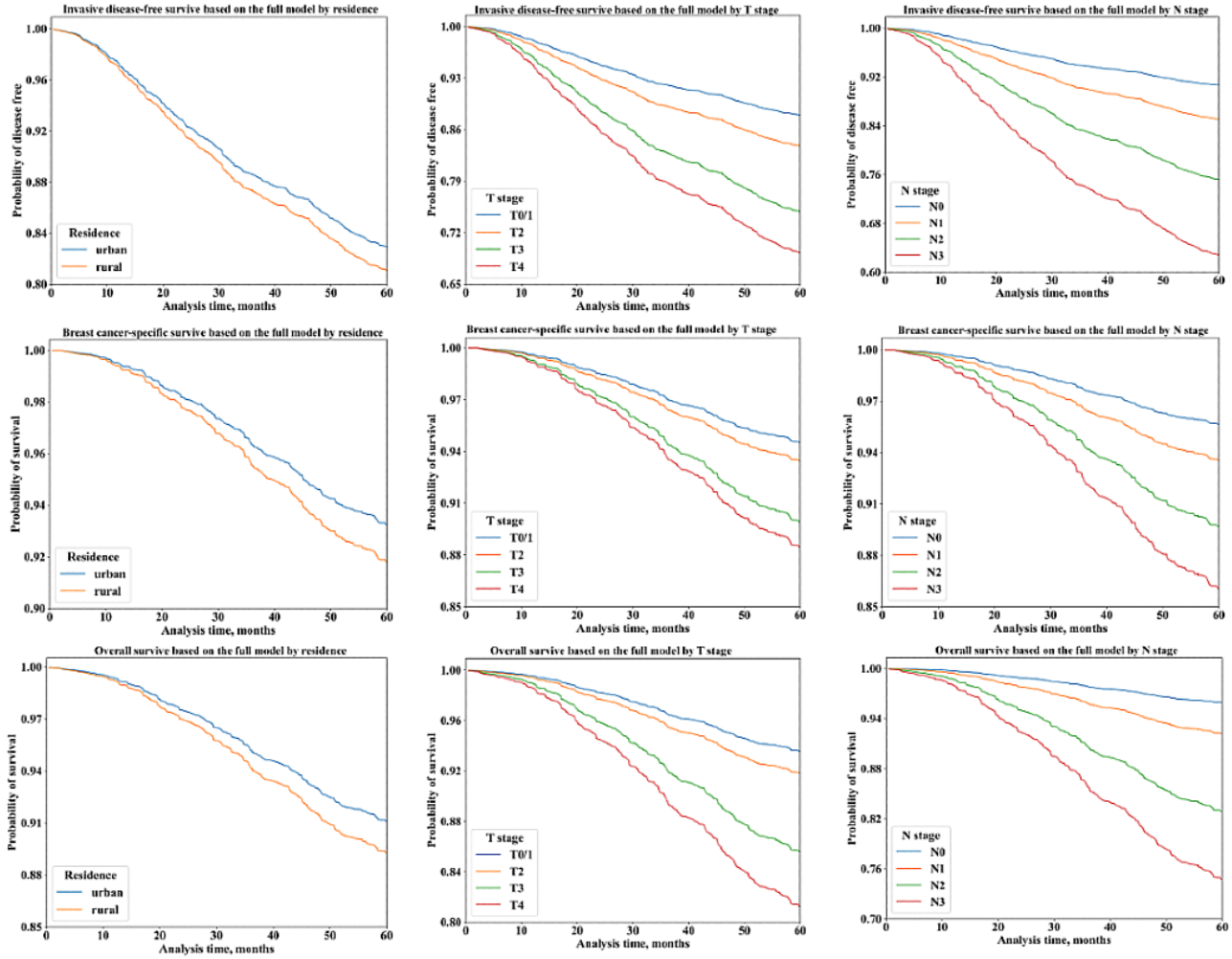
The training population included 3276 women with a median follow-up period of 7.82 (range 0.01-19.08) years. Of these,

499 women showed disease progression, 202 died from breast cancer, and 261 died from all causes within the first 5 years of follow-up. The test population included 1405 women with a median follow-up period of 8.00 (range 0.01-19.94) years. Of these, 211 women showed disease progression, 94 died from breast cancer, and 129 died from all causes within the first 5 years of follow-up. The validation population included 612 women with a median follow-up period of 5.16 (range 0.01-6.25) years. Of these, 109 women showed disease progression, 33 died from breast cancer, and 37 died from all causes within the first 5 years of follow-up. The demographic, tumor, and treatment characteristics for training, test, and validation data sets are described in Multimedia Appendix 2. The baseline data of patients in the training and test sets were similar, whereas several characteristics differed between training and validation data sets (Multimedia Appendix 2).

Prognostic Models Incorporating Demographic, Tumor, and Treatment Characteristics

Model development used baseline demographic, tumor, and treatment characteristics in the training data set. The full model included age at diagnosis, diagnosis year, menopausal status at diagnosis, residence, T stage, N stage, histological grade, receptor type (ER, PR, HER2), Ki67, surgery, chemotherapy regimens and adherence, radiotherapy, endocrine therapy and regimens. Figure 2 shows variable importance of each outcome according to the SHA*P* value plot. N stage, T stage, endocrine therapy, and radiotherapy ranked as the top features for patient outcomes. The partial dependence plot showed the contribution of a category for each feature (Multimedia Appendix 3-5). The survival curve for the full model based on selected factors is shown in Figure 3.

Figure 3. Invasive disease free survival based on the full model (A) by residence, (B) by T stage, and (C) by N stage. Breast cancer–specific survival based on the full model (D) by residence, (E) by T stage, and (F) by N stage. Overall survival based on the full model (G) by residence, (H) by T stage, and (I) by N stage.



Compared with the other models, the full model exhibited better AUROC with the training data set (disease progression: AUROC 0.76; cancer-specific mortality: AUROC 0.88; all-cause mortality: AUROC 0.82) (Figure 4). The cut-off points were 0.126, 0.064, and 0.072 for disease progression, cancer-specific mortality, and all-cause mortality, respectively. The full model also showed a better AUROC than those of the other models with the test data set (disease progression: AUROC 0.79; cancer-specific mortality: AUROC 0.80; all-cause mortality: AUROC 0.83), except for models B and C for cancer-specific mortality and model C for all-cause mortality (Figure 4). With the validation data set, the full model showed AUROC values comparable with those of the other models (disease progression: AUROC 0.79; cancer-specific mortality: AUROC 0.84; all-cause mortality: AUROC 0.88), except for an improved AUROC for cancer-specific mortality over the AUROC of model B (Figure

4). We also observed good model calibration for each model, except for disease progression prediction with the validation data set (Table 1 and Multimedia Appendix 10).

Figure 4. Discriminatory accuracy for predicting breast cancer outcomes: (A) disease progression (training), (B) disease progression (test), (C) disease progression (validation), (D) cancer-specific mortality (training), (E) cancer-specific mortality (test), (F) cancer-specific mortality (validation), (G) all-specific mortality (training), (H) all-specific mortality (test), and (I) all-specific mortality (validation). AUC: area under the curve.

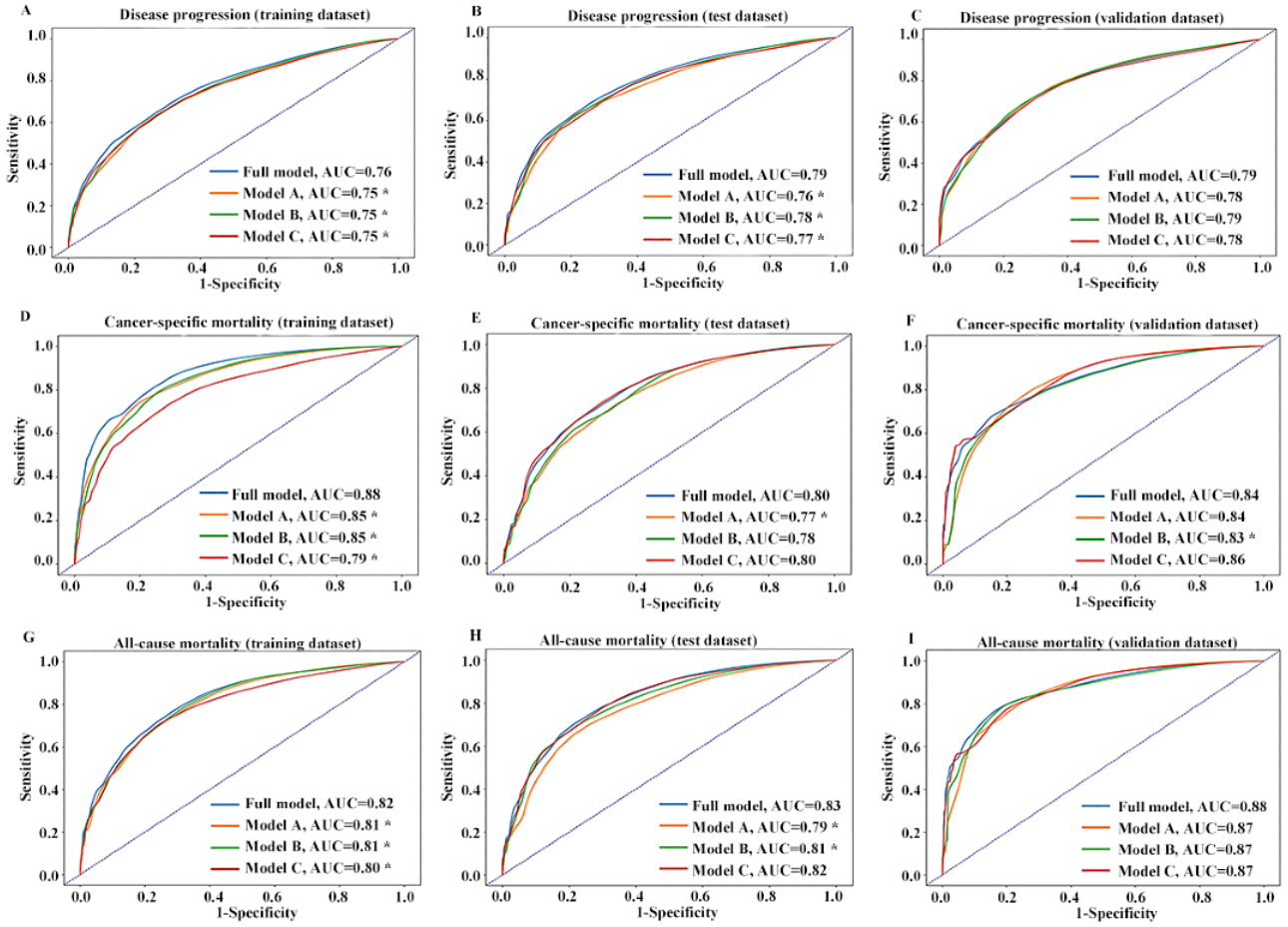


Table 1. Observed and predicted 5-year events.

Data set	Observed	Full model	*P* value	Model A	*P* value	Model B	*P* value	Model C	*P* value
Disease progression									
Test data	211	224.31	.35	221.03	.48	222.29	.43	220.37	.52
Validation data	109	88.05	.02	89.64	.03	89.16	.03	82.92	.02
Cancer-specific mortality									
Test data	94	98.42	.68	93.51	>.999	94.25	>.999	94.35	>.999
Validation data	33	36.03	.66	36.98	.55	35.53	.73	34.09	>.999
All-cause mortality									
Test data	129	122.18	.55	119.36	.38	118.03	.31	120.12	.42
Validation data	37	42.60	.42	43.22	.37	39.54	.74	41.26	.54

Subgroup Analyses

Discrimination of the full model with the test and validation data sets was evaluated using demographic and tumor characteristics (Table 2). The full model showed good discrimination in most subgroups of the test data set (AUROC 0.70-0.87), except in the ER–/PR–/HER2– and hormone receptor (HR)+/HER2+ subgroups for disease progression and cancer-specific mortality (AUROC 0.63-0.69). With the validation data set, the full model showed good AUROC values for all subgroups (AUROC 0.70-0.97). In addition, the full model was well calibrated in most subgroups of the test data set, except for underestimating the risk of all-cause mortality in the >64-year-old subgroup (*P*=.04) (Table 3). It also showed good calibration in most subgroups of the validation data set, except for underestimating the risk of cancer-specific mortality

of ER–/PR–/HER2– patients (4.65 events vs 11 events, *P*=.004) and underestimating the risk of disease progression of the 45 to 54-year-old (25.95 vs 39, *P*=.01), urban (58.91 vs 77, *P*=.01), and HR+/HER2+ (18.34 vs 33, *P*<.001) subgroups.

Table 2. AUROC by subgroup analysis.

Subgroup	Test data set, AUROC (95% CI)			Validation data set, AUROC (95% CI)		
	Disease progression	Cancer-specific mortality	All-cause mortality	Disease progression	Cancer-specific mortality	All-cause mortality
Age at diagnosis						
<45 years	0.79 (0.74-0.85)	0.79 (0.71-0.87)	0.83 (0.77-0.89)	0.80 (0.73-0.88)	0.91 (0.81-1.00)	0.94 (0.88-1.00)
45-54 years	0.80 (0.74-0.86)	0.79 (0.71-0.88)	0.81 (0.74-0.89)	0.79 (0.70-0.88)	0.84 (0.72-0.96)	0.85 (0.73-0.98)
55-64 years	0.75 (0.67-0.83)	0.80 (0.72-0.88)	0.82 (0.75-0.90)	0.77 (0.63-0.90)	0.80 (0.59-1.00)	0.83 (0.64-1.00)
>64 years	0.79 (0.67-0.92)	0.82 (0.66-0.97)	0.84 (0.74-0.93)	0.79 (0.65-0.94)	0.80 (0.60-1.00)	0.85 (0.73-0.98)
Residence						
Urban	0.78 (0.73-0.82)	0.80 (0.75-0.86)	0.82 (0.78-0.86)	0.78 (0.74-0.84)	0.84 (0.75-0.93)	0.90 (0.85-0.96)
Rural	0.81 (0.75-0.87)	0.77 (0.67-0.87)	0.84 (0.77-0.91)	0.80 (0.71-0.90)	0.84 (0.71-0.97)	0.84 (0.70-0.98)
Receptor type						
ER[a]–/PR[b]–/HER2[c]–	0.69 (0.61-0.78)	0.63 (0.52-0.74)	0.70 (0.60-0.79)	0.92 (0.83-1.00)	0.96 (0.92-1.00)	0.97 (0.94-1.00)
ER–/PR–/HER2+	0.75 (0.63-0.87)	0.86 (0.77-0.96)	0.85 (0.76-0.94)	0.70 (0.53-0.86)	0.87 (0.62-1.00)	0.87 (0.61-1.00)
HR[d]+/HER2-	0.84 (0.79-0.89)	0.84 (0.78-0.91)	0.87 (0.82-0.92)	0.73 (0.63-0.83)	0.75 (0.59-0.92)	0.83 (0.70-0.97)
HR+/HER2+	0.69 (0.55-0.82)	0.69 (0.48-0.89)	0.78 (0.63-0.94)	0.87 (0.81-0.94)	0.81 (0.65-0.97)	0.84 (0.70-0.97)

[a]ER: estrogen receptor.

[b]PR: progesterone receptor.

[c]HER2: human epidermal growth factor receptor.

[d]HR: hormone receptor.

Table 3. Observed and predicted 5-year events by subgroup analysis.

Data set and subgroup	Disease progression			Cancer-specific mortality			All-cause mortality		
	Observed	Predicted	*P* value	Observed	Predicted	*P* value	Observed	Predicted	*P* value
Test data set									
Age at diagnosis									
<45 years	71	78.61	.38	29	35.36	.31	38	40.09	.79
45-54 years	73	73.64	.99	27	33.63	.27	41	41.64	.98
55-64 years	46	52.81	.34	27	21.6	.27	32	29.4	.68
>64 years	21	19.26	.76	11	7.83	.32	18	11.05	.04
Residence									
Urban	155	170.96	.20	69	72.75	.69	93	89.95	.78
Rural	56	53.36	.74	25	25.67	.97	36	32.22	.54
Receptor type									
ER[a]–/PR[b]–/HER2[c]–	53	56.11	.69	29	24.42	.38	38	32.35	.33
ER–/PR–/HER2+	25	30.48	.30	10	14.48	.27	15	18.95	.39
HR+/HER2–	93	102.66	.33	39	43.06	.58	54	53.14	.96
HR+/ HER2+	23	18.02	.26	8	9.3	.79	12	9.17	.42
Validation data set									
Age at diagnosis									
<45 years	40	31.83	.14	7	12.86	.12	10	13.95	.34
45-54 years	39	25.95	.01	13	11.72	.81	13	13.38	>.999
55-64 years	19	22.27	.52	9	8.68	1.00	9	11.41	.55
>64 years	11	8	.33	4	2.77	.65	5	3.85	.73
Residence									
Urban	77	58.91	.01	18	23.6	.28	22	27.38	.33
Rural	32	29.14	.63	15	12.43	.54	15	15.22	>.999
Receptor type									
ER–/PR–/HER2–	15	10.72	.20	11	4.65	.004	11	6.57	.10
ER–/PR–/HER2+	15	13	.64	5	5.53	.99	5	6.9	.57
HR+/HER2–	35	32.02	.64	7	12.92	.12	9	14.3	.19
HR+/HER2+	33	18.34	<.001	7	7.88	.89	9	9.31	>.999

[a]ER: estrogen receptor.

[b]PR: progesterone receptor.

[c]HER2: human epidermal growth factor receptor.

Comparison with PREDICT

We also compared the performance of PREDICT with that of the full model. Both models showed good discrimination and similar AUROC values (0.78-0.84) with the test and validation data sets (Figure 5). However, based on our data, PREDICT was not well calibrated (Table 4). It overestimated the breast cancer specific (80.6 vs 27, 52.6 vs 19, $P<.001$) and all-cause mortalities (93.4 vs 39, 62.1 vs 21, $P<.001$), whereas our model exhibited good calibration.

Figure 5. Discriminatory accuracy for (A) cancer-specific mortality (test), (B) cancer-specific mortality (validation), (C) all-specific mortality (test), and (D) all-specific mortality (validation). AUC: area under the curve; WCH: West China Hospital.

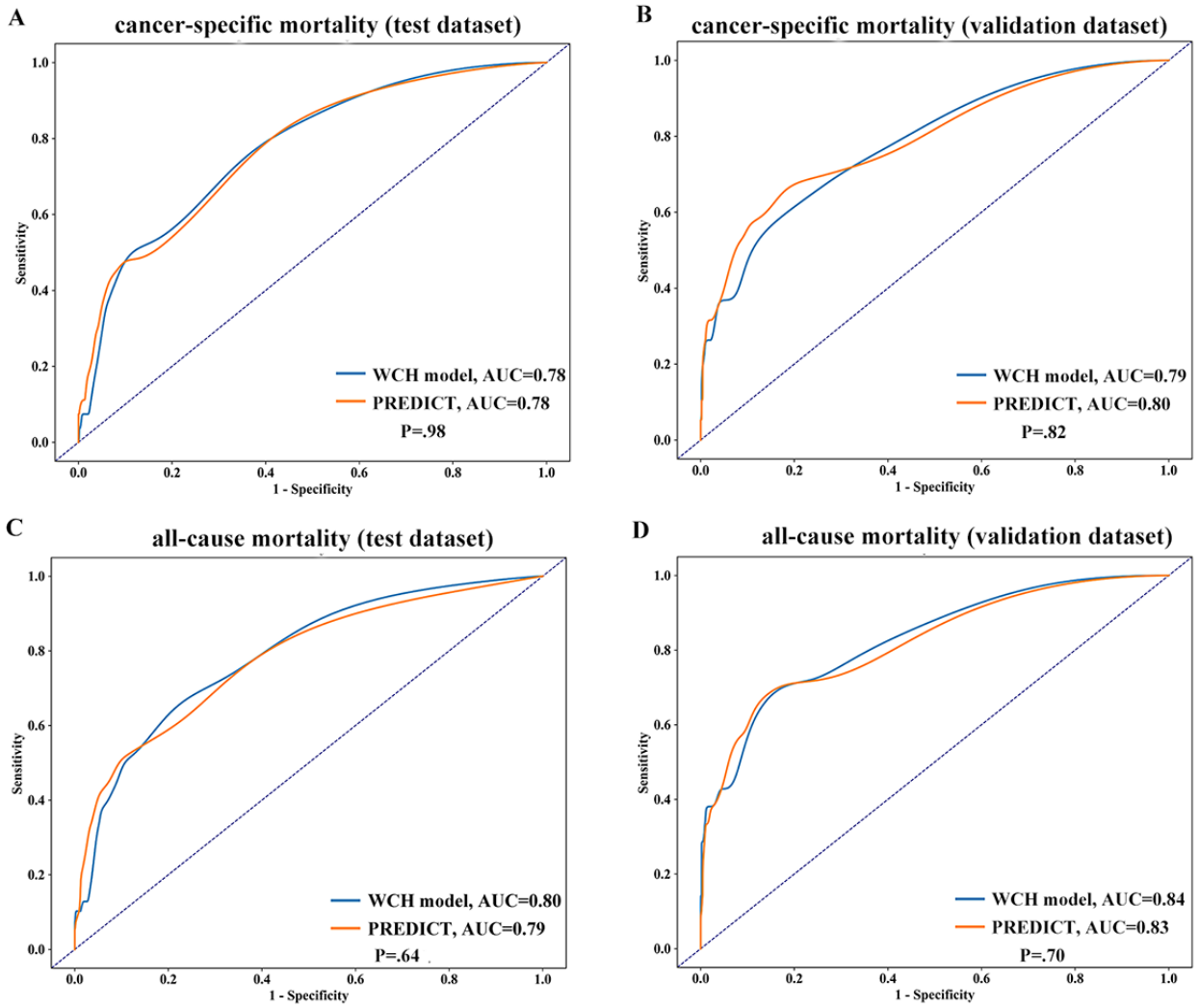


Table 4. Observed and predicted 5-year deaths by full model and PREDICT model.

Calibration	N	Breast cancer–specific mortality			All-cause mortality		
		Observed	Predicted	*P* value	Observed	Predicted	*P* value
Test data set							
PREDICT	602	27	80.6	<.001	39	93.4	<.001
West China Hospital model	602	27	35.1	.19	39	39.3	>.999
Validation data set							
PREDICT	486	19	52.6	<.001	21	62.1	<.001
West China Hospital model	486	19	25.2	.24	21	27.5	.24

Discussion

Principal Findings

Leveraging the real-world data of 5293 women with primary invasive early breast cancer, we developed a prognostic model to estimate the individual risk of disease progression, cancer-specific mortality, and all-cause mortality using machine learning. Good discriminatory accuracy and calibration were obtained by combining patient demographic, tumor, and treatment factors.

Adjuvant! Online and PREDICT are largely based on Caucasians and have been well validated and widely used in the United States and Western Europe [4,22,23]; however, several validation attempts in non-European countries and even in some European countries revealed suboptimal predictions [2,9-13]. Among the population composition of the race-specific

model developed by Wu et al [14], most patients were White, followed by Hispanic and African American, whereas only 518 patients were Asian. In this study, the full model was compared with 3 other models. Compared with model A (demographic and tumor variables) and model B (variables similar to those used in the published models), the full model (demographic, tumor, and treatment variables) exhibited better AUROC, indicating that the additional variables contributed to the improvement in the full model. However, the full model with XGBoost showed AUROC values comparable with those of model C (same variables using Cox proportional hazards regression) in the test and validation data sets, except for a significantly better AUROC for disease progression prediction with the test data set. This showed that the machine learning method, similar to the traditional method, may be suitable for constructing prognostic models based on survival data. There is increasing interest in applying machine learning to clinical data and offering personalized information to support clinical practice [24-27]. Moreover, machine learning provides an innovative approach to data analysis and imaging interpretation, which may be superior to conventional statistics [28]. The ability to automatically handle large multidimensional and multivariate data may ultimately reveal novel associations between specific features and important cancer outcomes. This helps to identify trends and patterns that would otherwise be obscure to investigators [29]. Therefore, a machine learning–based model may play an important role in patient risk stratification [30].

This study also compared the performance of PREDICT with that of our model and showed that the PREDICT algorithm overestimated mortality. This discrepancy is likely due to the lack of data on tumor detection methods [31] as well as to the lack of generalizability to the entire Chinese population. The validation of PREDICT based on an Asian population in another study revealed similar results [9], suggesting that attention should be paid to racial and ethnic differences [32]. Race-specific breast cancer prognosis models for White, Hispanic, and African American patients showed that racial disparity was evident in the distributions of several risk factors and the clinical presentation of the disease [14]. These results suggest that breast cancer prognostic model specific to the characteristics of different populations should be established. To the best of our knowledge, this is the first breast cancer prognosis model based on a Chinese population.

One major merit of our study was the large-scale prospective cohort design with virtually complete follow-up, largely limiting the common sources of bias. Although our study is based on a single institution, the large-scale cohort and complete coverage in West China Hospital guarantee the representativeness of breast cancer patients in Southwestern China. This study is based on real-world data recorded in the BCIMS. The BCIMS infrastructure ensured high quality data collection and virtually complete follow-up through regular interviews, which considerably restricted several common biases such as information and surveillance biases. Several studies have used real-world data to develop cancer models [33-38]. Real-world data are more representative of a patient's true state than clinical research data.

In real-word practices, some prognostic indicators were missing due to incomplete records of pathological diagnoses in early 2000s, such as histological grade and Ki67 percentage. Some HER2 status data were uncertain because HER2+ results obtained by immunohistochemistry were not further verified by fluorescence in situ hybridization. Although these missing data were inputted as unknown categories in the full model, the model's good performance relieved this concern to some extent. Moreover, the unknown categories were not related to patient outcome in model C by the Cox method.

The full model incorporated the residential status of breast cancer patients. The incidence of breast cancer in China is generally higher in urban than rural areas, but the associated mortality risk is considerably higher in rural areas [31]. Indeed, the residential status represents the socioeconomic status of Chinese patients to a large extent. Disparities exist between urban and rural patients in terms of lifestyle, medical insurance, ability to afford out-of-pocket treatment expenses, health service, geographical and travel issues, health education, and treatment intention and adherence [39,40]. These factors are associated with patient prognosis [39,41-45]. Moreover, with the progress of urbanization, the residential status of the population is undergoing dynamic changes and should be adjusted in future models.

Our study has some limitations. First, the proposed model showed poor AUROC values (0.63-0.69) for the ER–/PR–/HER2– and HR+/HER2+ subgroups in the test data set. However, it showed good AUROC values for these 2 subgroups in the validation data set (0.81-0.96), which relieves the concern. Notably, this difference in performance between the test and validation data sets was probably because the validation population was diagnosed and treated in 2013, with fewer instances of missing data. Second, the model did not include the variable of targeted therapy. Trastuzumab was approved in China in 2002, but because of its high cost and exclusion from reimbursement in Sichuan province until 2017, the number of HER2+ patients treated with trastuzumab was relatively small in our institution. Third, as a single-center study, our models were developed using a large-scale cohort in the training phase, and the test and validation groups were independent but from the same population. Therefore, validation in an external population is needed in the future.

Conclusions

We developed and validated a prognostic model for a Chinese population of patients with early-stage invasive breast cancer. Our model showed high discriminatory accuracy and good calibration, which may facilitate prognosis prediction and decision making in clinical practice for Chinese patients with breast cancer.

Acknowledgments

We thank Professor Paul Pharoah from the University of Cambridge for providing the PREDICT code and interpreting the validation results of PREDICT based on our data. The authors would like to thank AsiaEdit for providing linguistic assistance during the preparation of this manuscript. This work was supported by the key program of the Science and Technology Department of Sichuan Province (grant number: 2017SZ0005 to HZ), the 135 project for disciplines of excellence, West China Hospital, Sichuan University (grant number: ZYGD18012 to HB), and the Post-Doctor Research Project, West China Hospital, Sichuan University (grant number: 2019HXBH015 to LD).

Authors' Contributions

XZ, TL, JL, FY, DL, BF, and HZ conceived the study concept and design. PH, KH, YX, CL, DZ, TP, and XZ collected data. LD, PL, and DL performed statistical analysis. XZ and TL drafted the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The candidate features for model development.
[DOCX File , 61 KB - medinform_v8i11e19069_app1.docx]

Multimedia Appendix 2
Patients characteristics in the training, test, and validation data sets.
[DOCX File , 99 KB - medinform_v8i11e19069_app2.docx]

Multimedia Appendix 3
The contribution of predictors on disease progression in the full model.
[DOCX File , 893 KB - medinform_v8i11e19069_app3.docx]

Multimedia Appendix 4
The contribution of predictors on breast cancer mortality in the full model.
[DOCX File , 864 KB - medinform_v8i11e19069_app4.docx]

Multimedia Appendix 5
The contribution of predictors to all-cause mortality in the full model.
[DOCX File , 872 KB - medinform_v8i11e19069_app5.docx]

Multimedia Appendix 6
The importance of predictors in model A.
[DOCX File , 432 KB - medinform_v8i11e19069_app6.docx]

Multimedia Appendix 7
Supplementary methods.
[DOCX File , 79 KB - medinform_v8i11e19069_app7.docx]

Multimedia Appendix 8
The importance of predictors in model B.
[DOCX File , 514 KB - medinform_v8i11e19069_app8.docx]

Multimedia Appendix 9
Cox proportional hazards regression model for patient survival.
[DOCX File , 98 KB - medinform_v8i11e19069_app9.docx]

Multimedia Appendix 10
The calibration plot for each model.
[DOCX File , 543 KB - medinform_v8i11e19069_app10.docx]

References

1. Nicolini A, Ferrari P, Duffy MJ. Prognostic and predictive biomarkers in breast cancer: Past, present and future. Semin Cancer Biol 2018 Oct;52(Pt 1):56-73. [doi: 10.1016/j.semcancer.2017.08.010] [Medline: 28882552]
2. El Hage Chehade H, Wazir U, Mokbel K, Kasem A, Mokbel K. Do online prognostication tools represent a valid alternative to genomic profiling in the context of adjuvant treatment of early breast cancer? A systematic review of the literature. Am J Surg 2018 Jan;215(1):171-178. [doi: 10.1016/j.amjsurg.2017.05.006] [Medline: 28622841]
3. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001 Feb 15;19(4):980-991. [doi: 10.1200/JCO.2001.19.4.980] [Medline: 11181660]
4. Candido Dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res 2017 May 22;19(1):58 [FREE Full text] [doi: 10.1186/s13058-017-0852-3] [Medline: 28532503]
5. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, MINDACT Investigators. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med 2016 Aug 25;375(8):717-729. [doi: 10.1056/NEJMoa1602253] [Medline: 27557300]
6. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004 Dec 30;351(27):2817-2826. [doi: 10.1056/NEJMoa041588] [Medline: 15591335]
7. Dubsky P, Filipits M, Jakesz R, Rudas M, Singer CF, Greil R, Austrian BreastColorectal Cancer Study Group (ABCSG). EndoPredict improves the prognostic classification derived from common clinical guidelines in ER-positive, HER2-negative early breast cancer. Ann Oncol 2013 Mar;24(3):640-647 [FREE Full text] [doi: 10.1093/annonc/mds334] [Medline: 23035151]
8. Zhang Y, Schnabel CA, Schroeder BE, Jerevall P, Jankowitz RC, Fornander T, et al. Breast cancer index identifies early-stage estrogen receptor-positive breast cancer patients at risk for early- and late-distant recurrence. Clin Cancer Res 2013 Aug 01;19(15):4196-4205. [doi: 10.1158/1078-0432.CCR-13-0804] [Medline: 23757354]
9. Wong H, Subramaniam S, Alias Z, Taib NA, Ho G, Ng C, et al. The predictive accuracy of PREDICT: a personalized decision-making tool for Southeast Asian women with breast cancer. Medicine (Baltimore) 2015 Feb;94(8):e593 [FREE Full text] [doi: 10.1097/MD.0000000000000593] [Medline: 25715267]
10. Bhoo-Pathy N, Yip C, Hartman M, Saxena N, Taib NA, Ho G, et al. Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients. Eur J Cancer 2012 May;48(7):982-989 [FREE Full text] [doi: 10.1016/j.ejca.2012.01.034] [Medline: 22366561]
11. Quintyne KI, Woulfe B, Coffey JC, Gupta RK. Correlation between Nottingham Prognostic Index and Adjuvant! Online prognostic tools in patients with early-stage breast cancer in Mid-Western Ireland. Clin Breast Cancer 2013 Aug;13(4):233-238. [doi: 10.1016/j.clbc.2013.02.011] [Medline: 23829889]
12. Mook S, Schmidt MK, Rutgers EJ, van de Velde AO, Visser O, Rutgers SM, et al. Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study. Lancet Oncol 2009 Nov;10(11):1070-1076. [doi: 10.1016/S1470-2045(09)70254-2] [Medline: 19801202]
13. van Maaren MC, van Steenbeek CD, Pharoah PDP, Witteveen A, Sonke GS, Strobbe LJA, et al. Validation of the online prediction tool PREDICT v. 2.0 in the Dutch breast cancer population. Eur J Cancer 2017 Nov;86:364-372. [doi: 10.1016/j.ejca.2017.09.031] [Medline: 29100191]
14. Wu X, Ye Y, Barcenas CH, Chow W, Meng QH, Chavez-MacGregor M, et al. Personalized Prognostic Prediction Models for Breast Cancer Recurrence and Survival Incorporating Multidimensional Data. J Natl Cancer Inst 2017 Jul 01;109(7) [FREE Full text] [doi: 10.1093/jnci/djw314] [Medline: 28376179]
15. 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 08;8(6):e17364 [FREE Full text] [doi: 10.2196/17364] [Medline: 32510459]
16. Peng Z, Wei J, Lu X, Zheng H, Zhong X, Gao W, et al. Treatment and survival patterns of Chinese patients diagnosed with breast cancer between 2005 and 2009 in Southwest China: An observational, population-based cohort study. Medicine (Baltimore) 2016 Jun;95(25):e3865 [FREE Full text] [doi: 10.1097/MD.0000000000003865] [Medline: 27336872]
17. Fu B, Liu P, Lin J, Deng L, Hu K, Zheng H. Predicting Invasive Disease-Free Survival for Early-stage Breast Cancer Patients Using Follow-up Clinical Data. IEEE Trans Biomed Eng 2018 Nov 22. [doi: 10.1109/TBME.2018.2882867] [Medline: 30475709]
18. Yang L, Pelckmans K. Machine Learning Approaches to Survival Analysis: Case Studies in Microarray for Breast Cancer. IJMLC 2014;4(6):483-490. [doi: 10.7763/ijmlc.2014.v6.459]
19. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. 2005 Presented at: International Conference on Machine Learning; August 7; Bonn, Germany p. 625-632. [doi: 10.1145/1102351.1102430]
20. Lundberg S, Lee S. A Unified Approach to Interpreting Model Predictions. USA: Curran Associates Inc; 2017 Presented at: 31st Conference on Neural Information Processing Systems; 2017; Long Beach, CA.

21. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics 2015 Mar 31;24(1):44-65. [doi: 10.1080/10618600.2014.907095]
22. Campbell HE, Taylor MA, Harris AL, Gray AM. An investigation into the performance of the Adjuvant! Online prognostic programme in early breast cancer for a cohort of patients in the United Kingdom. Br J Cancer 2009 Oct 06;101(7):1074-1084 [FREE Full text] [doi: 10.1038/sj.bjc.6605283] [Medline: 19724274]
23. Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, Norris BD, et al. Population-based validation of the prognostic model ADJUVANT! for early breast cancer. J Clin Oncol 2005 Apr 20;23(12):2716-2725. [doi: 10.1200/JCO.2005.06.178] [Medline: 15837986]
24. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing Machine Learning in Radiology Practice and Research. AJR Am J Roentgenol 2017 Apr;208(4):754-760. [doi: 10.2214/AJR.16.17224] [Medline: 28125274]
25. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision. Radiology 2018 Mar;286(3):810-818. [doi: 10.1148/radiol.2017170549] [Medline: 29039725]
26. Saha A, Harowicz MR, Wang W, Mazurowski MA. A study of association of Oncotype DX recurrence score with DCE-MRI characteristics using multivariate machine learning models. J Cancer Res Clin Oncol 2018 May;144(5):799-807 [FREE Full text] [doi: 10.1007/s00432-018-2595-7] [Medline: 29427210]
27. Kim W, Kim KS, Lee JE, Noh D, Kim S, Jung YS, et al. Development of novel breast cancer recurrence prediction model using support vector machine. J Breast Cancer 2012 Jun;15(2):230-238 [FREE Full text] [doi: 10.4048/jbc.2012.15.2.230] [Medline: 22807942]
28. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol 2017 May 30;69(21):2657-2664. [doi: 10.1016/j.jacc.2017.03.571] [Medline: 28545640]
29. Hernandez-Suarez DF, Kim Y, Villablanca P, Gupta T, Wiley J, Nieves-Rodriguez BG, et al. Machine Learning Prediction Models for In-Hospital Mortality After Transcatheter Aortic Valve Replacement. JACC Cardiovasc Interv 2019 Jul 22;12(14):1328-1338. [doi: 10.1016/j.jcin.2019.06.013] [Medline: 31320027]
30. 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]
31. Fan L, Strasser-Weippl K, Li J, St LJ, Finkelstein DM, Yu K, et al. Breast cancer in China. Lancet Oncol 2014 Jun;15(7):e279-e289. [doi: 10.1016/S1470-2045(13)70567-9] [Medline: 24872111]
32. Iqbal J, Ginsburg O, Rochon PA, Sun P, Narod SA. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. JAMA 2015 Jan 13;313(2):165-173. [doi: 10.1001/jama.2014.17322] [Medline: 25585328]
33. Yao Z, Liao W, Ho C, Chen K, Shih J, Chen J, et al. Real-World Data on Prognostic Factors for Overall Survival in EGFR Mutation-Positive Advanced Non-Small Cell Lung Cancer Patients Treated with First-Line Gefitinib. Oncologist 2017 Sep;22(9):1075-1083 [FREE Full text] [doi: 10.1634/theoncologist.2016-0331] [Medline: 28507206]
34. Hang J, Wu L, Zhu L, Sun Z, Wang G, Pan J, et al. Prediction of overall survival for metastatic pancreatic cancer: Development and validation of a prognostic nomogram with data from open clinical trial and real-world study. Cancer Med 2018 Jun 01 [FREE Full text] [doi: 10.1002/cam4.1573] [Medline: 29856121]
35. Mandoj C, Pizzuti L, Sergi D, Sperduti I, Mazzotta M, Di Lauro L, et al. Observational study of coagulation activation in early breast cancer: development of a prognostic model based on data from the real world setting. J Transl Med 2018 May 16;16(1):129 [FREE Full text] [doi: 10.1186/s12967-018-1511-x] [Medline: 29769125]
36. Fernández Montes A, López López C, Argilés Martínez G, Páez López D, López Muñoz AM, García Paredes B, et al. Prognostic Nomogram and Patterns of Use of FOLFIRI-Aflibercept in Advanced Colorectal Cancer: A Real-World Data Analysis. Oncologist 2019 Aug;24(8):e687-e695 [FREE Full text] [doi: 10.1634/theoncologist.2018-0824] [Medline: 31147489]
37. Pobiruchin M, Bochum S, Martens UM, Kieser M, Schramm W. A method for using real world data in breast cancer modeling. J Biomed Inform 2016 Apr;60:385-394 [FREE Full text] [doi: 10.1016/j.jbi.2016.01.017] [Medline: 26854868]
38. Kim JJ, Tosteson AN, Zauber AG, Sprague BL, Stout NK, Alagoz O, Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) consortium. Cancer Models and Real-world Data: Better Together. J Natl Cancer Inst 2016 Feb;108(2) [FREE Full text] [doi: 10.1093/jnci/djv316] [Medline: 26538628]
39. Peng Z, Wei J, Lu X, Zheng H, Zhong X, Gao W, et al. Diagnosis and treatment pattern among rural and urban breast cancer patients in Southwest China from 2005 to 2009. Oncotarget 2016 Nov 22;7(47):78168-78179 [FREE Full text] [doi: 10.18632/oncotarget.11375] [Medline: 27556301]
40. Wang F, Yu L, Wang F, Liu L, Guo M, Gao D, et al. Risk factors for breast cancer in women residing in urban and rural areas of eastern China. J Int Med Res 2015 Dec;43(6):774-789. [doi: 10.1177/0300060515592901] [Medline: 26475794]
41. Hershman DL, Shao T, Kushi LH, Buono D, Tsai WY, Fehrenbacher L, et al. Early discontinuation and non-adherence to adjuvant hormonal therapy are associated with increased mortality in women with breast cancer. Breast Cancer Res Treat 2011 Apr;126(2):529-537 [FREE Full text] [doi: 10.1007/s10549-010-1132-4] [Medline: 20803066]

42. He W, Smedby KE, Fang F, Olsson H, Margolin S, Hall P, et al. Treatment Restarting After Discontinuation of Adjuvant Hormone Therapy in Breast Cancer Patients. J Natl Cancer Inst 2017 Oct 01;109(10). [doi: 10.1093/jnci/djx041] [Medline: 28423398]
43. Barron TI, Cahir C, Sharp L, Bennett K. A nested case-control study of adjuvant hormonal therapy persistence and compliance, and early breast cancer recurrence in women with stage I-III breast cancer. Br J Cancer 2013 Sep 17;109(6):1513-1521 [FREE Full text] [doi: 10.1038/bjc.2013.518] [Medline: 24002590]
44. Roder D, de Silva P, Zorbas HM, Kollias J, Malycha PL, Pyke CM, et al. Survival from breast cancer: an analysis of Australian data by surgeon case load, treatment centre location, and health insurance status. Aust Health Rev 2012 Aug;36(3):342-348. [doi: 10.1071/AH11060] [Medline: 22935129]
45. Hsu CD, Wang X, Habif DV, Ma CX, Johnson KJ. Breast cancer stage variation and survival in association with insurance status and sociodemographic factors in US women 18 to 64 years old. Cancer 2017 Aug 15;123(16):3125-3131 [FREE Full text] [doi: 10.1002/cncr.30722] [Medline: 28440864]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BCIMS: Breast Cancer Information Management System
ER: estrogen receptor
HR: hormone receptor
HER2: human epidermal growth factor receptor 2
PR: progesterone receptor
SHAP: Shapley additive explanations
XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 05.04.20; peer-reviewed by H Jiang, L Yongping; comments to author 26.07.20; revised version received 07.08.20; accepted 16.09.20; published 09.11.20.

Please cite as:
Zhong X, Luo T, Deng L, Liu P, Hu K, Lu D, Zheng D, Luo C, Xie Y, Li J, He P, Pu T, Ye F, Bu H, Fu B, Zheng H
Multidimensional Machine Learning Personalized Prognostic Model in an Early Invasive Breast Cancer Population-Based Cohort in China: Algorithm Validation Study
JMIR Med Inform 2020;8(11):e19069
URL: http://medinform.jmir.org/2020/11/e19069/
doi:10.2196/19069
PMID:33164899

Original Paper

Developing a Predictive Model for Asthma-Related Hospital Encounters in Patients With Asthma in a Large, Integrated Health Care System: Secondary Analysis

Gang Luo[1], DPhil; Claudia L Nau[2], DPhil; William W Crawford[3], MD; Michael Schatz[2,4], MSc, MD; Robert S Zeiger[2,4], MD, DPhil; Emily Rozema[2], MPH; Corinna Koebnick[2], DPhil

[1]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

[2]Department of Research & Evaluation, Kaiser Permanente Southern California, Pasadena, CA, United States

[3]Department of Allergy and Immunology, Kaiser Permanente South Bay Medical Center, Harbor City, CA, United States

[4]Department of Allergy, Kaiser Permanente Southern California, San Diego, CA, United States

Corresponding Author:
Gang Luo, DPhil
Department of Biomedical Informatics and Medical Education
University of Washington
UW Medicine South Lake Union, 850 Republican Street
Building C, Box 358047
Seattle, WA, 98195
United States
Phone: 1 206 221 4596
Fax: 1 206 221 2671
Email: gangluo@cs.wisc.edu

Abstract

Background: Asthma causes numerous hospital encounters annually, including emergency department visits and hospitalizations. To improve patient outcomes and reduce the number of these encounters, predictive models are widely used to prospectively pinpoint high-risk patients with asthma for preventive care via care management. However, previous models do not have adequate accuracy to achieve this goal well. Adopting the modeling guideline for checking extensive candidate features, we recently constructed a machine learning model on Intermountain Healthcare data to predict asthma-related hospital encounters in patients with asthma. Although this model is more accurate than the previous models, whether our modeling guideline is generalizable to other health care systems remains unknown.

Objective: This study aims to assess the generalizability of our modeling guideline to Kaiser Permanente Southern California (KPSC).

Methods: The patient cohort included a random sample of 70.00% (397,858/568,369) of patients with asthma who were enrolled in a KPSC health plan for any duration between 2015 and 2018. We produced a machine learning model via a secondary analysis of 987,506 KPSC data instances from 2012 to 2017 and by checking 337 candidate features to project asthma-related hospital encounters in the following 12-month period in patients with asthma.

Results: Our model reached an area under the receiver operating characteristic curve of 0.820. When the cutoff point for binary classification was placed at the top 10.00% (20,474/204,744) of patients with asthma having the largest predicted risk, our model achieved an accuracy of 90.08% (184,435/204,744), a sensitivity of 51.90% (2259/4353), and a specificity of 90.91% (182,176/200,391).

Conclusions: Our modeling guideline exhibited acceptable generalizability to KPSC and resulted in a model that is more accurate than those formerly built by others. After further enhancement, our model could be used to guide asthma care management.

International Registered Report Identifier (IRRID): RR2-10.2196/resprot.5039

(JMIR Med Inform 2020;8(11):e22689) doi:10.2196/22689

KEYWORDS

asthma; forecasting; machine learning; patient care management; risk factors

Introduction

Background

About 8.4% of people in the United States have asthma [1], which causes over 3000 deaths, around 500,000 hospitalizations, and over 2 million emergency department (ED) visits each year [1,2]. To improve patient outcomes and cut the number of asthma-related hospital encounters including ED visits and hospitalizations, predictive models are widely used to prospectively pinpoint high-risk patients with asthma for preventive care via care management. This is the case with health care systems such as the University of Washington Medicine, Kaiser Permanente Northern California [3], and Intermountain Healthcare, and with other health plans in 9 of 12 metropolitan communities [4]. Once a patient is identified as high risk and placed into a care management program, a care manager will call the patient periodically to assess asthma control, adjust asthma medications, and make appointments for needed care or testing. Successful care management can help patients with asthma obtain better outcomes, thereby avoiding up to 40% of their future hospital encounters [5-8].

A care management program has a limited service capacity and usually enrolls ≤3% of patients [9] with a given condition, which places a premium on enrolling at-risk patients. Therefore, the accuracy of the adopted predictive model (or lack thereof) puts an upper bound on the effectiveness of the program. Previously, several researchers have developed several models for projecting asthma-related hospital encounters in patients with asthma [3,10-22]. Each of these models would consider only a few features, miss more than half of patients who will have future asthma-related hospital encounters, and incorrectly project future asthma-related hospital encounters for many other patients with asthma [23]. These errors lead to suboptimal patient outcomes, including hospital encounters and unnecessary health care costs because of unneeded care management program enrollment. When building machine learning models on nonmedical data, people often follow the modeling guideline of checking extensive candidate features to boost model accuracy [24-27]. Adopting this modeling guideline to the medical domain, we recently constructed a machine learning model on Intermountain Healthcare data to project asthma-related hospital encounters in the following 12-month period in patients with asthma [23]. Compared with previous models, our model boosts the area under the receiver operating characteristic curve (AUC) by at least 0.049 to 0.859. Although this is encouraging, it remains unknown whether our modeling guideline is generalizable to other health care systems.

Objectives

This study aims to assess the generalizability of our modeling guideline to Kaiser Permanente Southern California (KPSC). Similar to our Intermountain Healthcare model [23], our KPSC model uses administrative and clinical data to project asthma-related hospital encounters (ED visits and hospitalizations) in patients with asthma. The categorical dependent variable has 2 possible values—whether the patient with asthma will have asthma-related hospital encounters in the following 12-month period or not. This study describes the construction and evaluation of our KPSC model.

Methods

The methods adopted in this study are similar to those used in our previous paper [23].

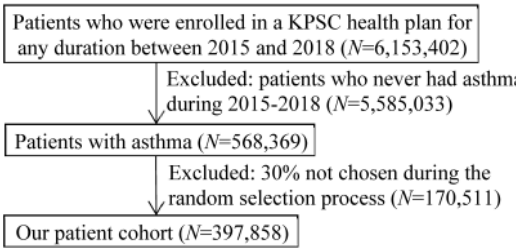
Ethics Approval and Study Design

In this study, we performed a secondary analysis of computerized administrative and clinical data. This study was approved by the institutional review boards of the University of Washington Medicine and KPSC.

Patient Population

As shown in Figure 1, our patient cohort was based on patients with asthma who were enrolled in a KPSC health plan for any duration between 2015 and 2018. Owing to internal regulatory processes, the patient cohort was restricted to a random sample of 70.00% (397,858/568,369) of eligible patients. This sample size is the maximum that KPSC allows for sharing its data with an institution outside of Kaiser Permanente for research. As the largest integrated health care system in Southern California with 227 clinics and 15 hospitals, KPSC offers care to approximately 19% of Southern California residents [28]. A patient was deemed to have asthma in a particular year if the patient had one or more diagnosis codes of asthma (International Classification of Diseases [ICD], Tenth Revision [ICD-10]: J45.x; ICD, Ninth Revision [ICD-9]: 493.0x, 493.1x, 493.8x, 493.9x) recorded in the encounter billing database in that year [11,29,30]. The exclusion criterion was that the patient died during that year. If a patient had no diagnosis code of asthma in any subsequent year, the patient was deemed to have no asthma in that subsequent year.

Figure 1. The patient cohort selection process. KPSC: Kaiser Permanente Southern California.



Prediction Target (the Dependent Variable)

For each patient identified as having asthma in a particular year, the outcome was whether the patient had any asthma-related hospital encounter in the following year. An asthma-related hospital encounter is an ED visit or hospitalization with asthma as the principal diagnosis (ICD-10: J45.x; ICD-9: 493.0x, 493.1x, 493.8x, 493.9x). For every patient with asthma, the patient's data up to the end of every calendar year were used to project the patient's outcome in the following year as long as the patient was deemed to have asthma in the previous year and was also enrolled in a KPSC health plan at the end of the previous year.

Data Set

For the patients in our patient cohort, we used their entire electronically available patient history at KPSC. At KPSC, various kinds of information on its patients has been recorded in the electronic medical record system since 2010. In addition, we had electronic records of the patients' diagnosis codes starting from 1981, regardless of whether they were stored in the electronic medical record system. From the research data warehouse at KPSC, we retrieved an administrative and clinical data set, including information regarding our patient cohort's encounters and medication dispensing at KPSC from 2010 to 2018 and diagnosis codes at KPSC from 1981 to 2018. Owing to regulatory and privacy concerns, the data set is not publicly available.

Features (Independent Variables)

We examined 2 types of candidate features—basic and extended. A basic feature and its corresponding extended features differ only in the year of the data used for feature computation. We considered 307 basic candidate features listed in Multimedia Appendix 1 [31]. Covering a wide range of characteristics, these basic candidate features were computed from the structured attributes in our data set. In Multimedia Appendix 1, unless the word *different* shows up, every mention of the number of a given type of item such as medications counts multiplicity. As defined in our previous paper [23], major visits for asthma include ED visits and hospitalizations with an asthma diagnosis code and outpatient visits with a primary diagnosis of asthma. Outpatient visits with a secondary but no primary diagnosis of asthma is regarded as minor visits for asthma.

Every input data instance to the model targets a unique (patient, index year) pair and is employed to forecast the patient's outcome in the following year. For the (patient, index year) pair, the patient's primary care provider (PCP), age, and home address were computed as of the end of the index year. The basic candidate features of history of bronchiolitis, the number of years since the first asthma-coded encounter in the data set, premature birth, family history of asthma, and the number of years since the first encounter for chronic obstructive pulmonary disease in the data set were computed using the data from 1981 to the index year. All of the allergy features and the features derived from the problem list were computed using the data from 2010 to the index year. One basic candidate feature was computed using the data in the index and preindex years: the proportion of patients who had asthma-related hospital encounters in the index year out of all of the patients of the patient's PCP with asthma in the preindex year. The other 277 basic candidate features were computed using the data in the index year.

In addition to the basic candidate features, we also checked extended candidate features. Our Intermountain Healthcare model [23] was built using the extreme gradient boosting (XGBoost) machine learning classification algorithm [32]. As detailed in Hastie et al [33], XGBoost automatically computes the importance value of every feature as the fractional contribution of the feature to the model. Previously, we showed that ignoring those features with importance values <0.01 led to a little drop in model accuracy [23]. Using the basic candidate features and the model construction method described below, we built an initial XGBoost model on KPSC data. As a patient's demographic features rarely change over time, no extended candidate feature was formed for any of the basic demographic features. For each basic candidate feature that was nondemographic, was computed on the data in the index year, and had an importance value 0.01 in the initial XGBoost model, we computed 2 related extended candidate features, one using the data in the preindex year and another using the data in the year that was 2 years before the index year. The only difference between the extended candidate features and the basic feature is the year of the data used for feature computation. For instance, for the basic candidate feature *number of ED visits in 2016*, the 2 related extended candidate features are the number of ED visits in 2015 and the number of ED visits in 2014. In brief, we formed extended candidate features for only those suitable and important basic candidate features. Our intuition is that among all possible ones that could be formed, these extended candidate features are most promising with regard to additional predictive power. For the other basic candidate features with lower importance values, those extended candidate features that could possibly be formed for them tend to have little extra predictive power and can be ignored. Given the finite data instances available for model training, this feature extending approach avoids a large rise in the number of candidate features, which may cause sample size issues. We considered all of the basic and extended candidate features when building our final predictive model.

Data Analysis

Data Preparation

Peak expiratory flow values are available in our KPSC data set but not in the Intermountain Healthcare data set used in our previous paper [23]. On the basis of the upper and lower bounds given by a medical expert (MS) in our team, all peak expiratory flow values >700 were regarded as biologically implausible. Using this criterion and the same data preparation method adopted in our previous paper [23], we normalized data, identified biologically implausible values, and set them to missing. As the outcomes were from the following year and the extended candidate features were computed using the data from up to 2 years before the index year, our data set contained 6 years of effective data (2012-2017) over a total of 9 years (2010-2018). In clinical practice, a model is trained on historical data and then applied to future years' data. To mirror this, the

2012 to 2016 data were used as the training set for model training. The 2017 data were employed as the test set to gauge model performance.

Performance Metrics

As shown in the formulas below and Table 1, we adopted 6 standard metrics to assess model performance: accuracy, specificity, sensitivity, negative predictive value (NPV), positive predictive value (PPV), and AUC.

Accuracy=(TP+TN)/(TP+TN+FP+FN),

Specificity=TN/(TN+FP),

Sensitivity=TP/(TP+FN),

Negative predictive value=TN/(TN+FN),

Positive predictive value=TP/(TP+FP).

We performed a 1000-fold bootstrap analysis [34] to compute the 95% CIs of these performance measures. We plotted the receiver operating characteristic (ROC) curve to show the tradeoff between sensitivity and specificity.

Table 1. The error matrix.

Outcome class	Asthma-related hospital encounters in the following year	No asthma-related hospital encounter in the following year
Projected asthma-related hospital encounters in the following year	TP[a]	FP[b]
Projected no asthma-related hospital encounter in the following year	FN[c]	TN[d]

[a]TP: true positive.

[b]FP: false positive.

[c]FN: false negative.

[d]TN: true negative.

Classification Algorithms

We employed Waikato Environment for Knowledge Analysis (WEKA) Version 3.9 [35] to build machine learning models. As a major open source toolkit for machine learning and data mining, WEKA integrates many classic feature selection techniques and machine learning algorithms. We examined the 39 native machine learning classification algorithms in WEKA, as shown in the web-based appendix of our previous paper [23] and the XGBoost classification algorithm [32] realized in the XGBoost4J package [36]. As an ensemble of decision trees, XGBoost implements gradient boosting in a scalable and efficient manner. As XGBoost takes only numerical features as its inputs, we converted every categorical feature to one or more binary features through one-hot encoding before giving the feature to XGBoost. We employed our previously developed automatic and efficient machine learning model selection method [37] and the 2012 to 2016 training data to automatically choose, among all of the applicable ones, the classification algorithm, feature selection technique, hyperparameter values, and data balancing method for managing imbalanced data. On average, our method runs 28 times faster and achieves an 11% lower model error rate than the Auto-WEKA automatic model selection method [37,38].

Assessing the Generalizability of our Intermountain Healthcare Model to KPSC

This study mainly assessed our modeling guideline's generalizability to KPSC by using the KPSC training set to train several models and assessing their performance on the KPSC test set. In addition, we assessed our Intermountain Healthcare model's [23] generalizability to KPSC. Using the Intermountain Healthcare data set and the top 21 features with an importance value computed by XGBoost ≥0.01, we formerly built a simplified Intermountain Healthcare model [23]. The simplified model retained almost all of the predictive power of our full Intermountain Healthcare model. Our KPSC data set included these 21 features but not all of the 142 features used in our full Intermountain Healthcare model. We assessed our simplified Intermountain Healthcare model's performance on the KPSC test set twice, once after retraining the model on the KPSC training set and once using the model trained on the Intermountain Healthcare data set without retraining the model on the KPSC training set.

Results

Clinical and Demographic Characteristics of the Patient Cohorts

Every data instance targets a unique (patient, index year) pair. Multimedia Appendix 1 displays the clinical and demographic characteristics of our patient cohort during the time periods of 2012 to 2016 and 2017. The set of characteristics during 2012 to 2016 is similar to that during 2017. During 2012 to 2016 and 2017, 2.42% (18,925/782,762) and 2.13% (4353/204,744) of data instances were associated with asthma-related hospital encounters in the following year, respectively.

Table 2 shows for each clinical or demographic characteristic, the statistical test results on whether the data instances linking to future asthma-related hospital encounters and those linking to no future asthma-related hospital encounter had the same distribution. These 2 sets of data instances had the same distribution when the *P* value is ≥.05, and distinct distributions when the *P* value is <.05. In Table 2, all of the *P* values <.05 are marked in italics.

Table 2. For each clinical or demographic characteristic, the statistical test results on whether the data instances linking to future asthma-related hospital encounters and those linking to no future asthma-related hospital encounter had the same distribution.

Characteristics	*P* value for the 2012-2016 data	*P* value for the 2017 data
Age (years)	*<.001* [a,b]	*<.001* [a]
Gender	*<.001* [c]	*.01* [c]
Race	*<.001* [c]	*<.001* [c]
Ethnicity	*<.001* [c]	*<.001* [c]
Insurance category	*<.001* [c]	*<.001* [c]
Number of years since the first asthma-coded encounter in the data set	.78[a]	*.006* [a]
Asthma medication fill		
Inhaled corticosteroid	*<.001* [c]	*<.001* [c]
Inhaled corticosteroid and long-acting beta-2 agonist combination	*<.001* [c]	*<.001* [c]
Leukotriene modifier	*<.001* [c]	*<.001* [c]
Long-acting beta-2 agonist	*<.001* [c]	*<.001* [c]
Mast cell stabilizer	>.99[c]	>.99[c]
Short-acting, inhaled beta-2 agonist	*<.001* [c]	*<.001* [c]
Systemic corticosteroid	*<.001* [c]	*<.001* [c]
Comorbidity		
Allergic rhinitis	*<.001* [c]	*<.001* [c]
Anxiety or depression	*<.001* [c]	*<.001* [c]
Bronchopulmonary dysplasia	*<.001* [c]	>.99[c]
Chronic obstructive pulmonary disease	*<.001* [c]	*<.001* [c]
Cystic fibrosis	>.99[c]	.52[c]
Eczema	*<.001* [c]	*<.001* [c]
Gastroesophageal reflux	*<.001* [c]	*<.001* [c]
Obesity	*<.001* [c]	*<.001* [c]
Premature birth	*<.001* [c]	*<.001* [c]
Sinusitis	.33[c]	.06[c]
Sleep apnea	*.003* [c]	*<.001* [c]
Smoking status	*<.001* [c]	*<.001* [c]

[a]*P* values obtained by performing the Cochran-Armitage trend test [39].

[b]*P* values <.05 marked in italics.

[c]*P* values obtained by performing the chi-square two-sample test.

Classification Algorithm and Features Used

Before building our final model, the importance values of the basic candidate features were computed once on our initial XGBoost model. This led to us examining 30 extended candidate features in addition to the 307 basic candidate features. With these 337 basic and extended candidates features as inputs, our automatic model selection method [37] picked the XGBoost classification algorithm [32]. As an ensemble of decision trees, XGBoost can handle missing feature values naturally. Our final predictive model was built using XGBoost, and the 221 features shown in descending order of importance value in Multimedia Appendix 1. The other features had no additional predictive power and were automatically dropped by XGBoost.

Performance Measures of the Final KPSC Model

On the KPSC test set, our final model achieved an AUC of 0.820 (95% CI 0.813-0.826). Figure 2 displays the ROC curve

of our final model. Table 3 displays the performance measures of our final model when various top percentages of patients having the largest predicted risk were adopted as the cutoff point for performing binary classification. When this percentage was at 10.00% (20,474/204,744), our final model achieved an accuracy of 90.08% (184,435/204,744; 95% CI 89.95-90.21), a sensitivity of 51.90% (2259/4353; 95% CI 50.44-53.42), a specificity of 90.91% (182,176/200,391; 95% CI 90.78-91.03), a PPV of 11.03% (2259/20,474; 95% CI 10.59-11.46), and an NPV of 98.86% (182,176/184,270; 95% CI 98.81-98.91). Table 4 gives the corresponding error matrix of our final model.

When we excluded the extended candidate features and considered only the basic candidate features, the AUC of our model dropped to 0.809. Several basic candidate features, such as the number of years since the first asthma-coded encounter in the data set, needed over one year of past data to calculate. When we further excluded these multiyear candidate features and considered only those basic candidate features calculated on 1 year of past data, the model's AUC dropped to 0.807.

Without precluding any feature from being considered, the model trained on data from both children (aged <18 years) with asthma and adults (aged ≥18 years) with asthma gained an AUC of 0.815 in children with asthma and an AUC of 0.817 in adults with asthma. In comparison, the model trained only on data from children with asthma gained an AUC of 0.811 in children with asthma. The model trained only on data from adults with asthma gained an AUC of 0.818 in adults with asthma.

If we adopted only the top 25 features shown in Multimedia Appendix 1 with an importance value ≥0.01 and removed the other 312 features, the model's AUC dropped from 0.820 to 0.800 (95% CI 0.793-0.808). When the top 10.00% (20,474/204,744) of patients having the largest predicted risk were adopted as the cutoff point for doing binary classification, the model's accuracy dropped from 90.08% (184,435/204,744) to 89.96% (184,185/204,744; 95% CI 89.83-90.08), sensitivity dropped from 51.90% (2259/4353) to 49.02% (2134/4353; 95% CI 47.71-50.55), specificity dropped from 90.91% (182,176/200,391) to 90.85% (182,051/200,391; 95% CI 90.72-90.97), PPV dropped from 11.03% (2259/20,474) to 10.42% (2134/20,474; 95% CI 10.03-10.86), and NPV dropped from 98.86% (182,176/184,270) to 98.80% (182,051/184,270; 95% CI 98.75-98.85).

Figure 2. The receiver operating characteristic curve of our final predictive model.

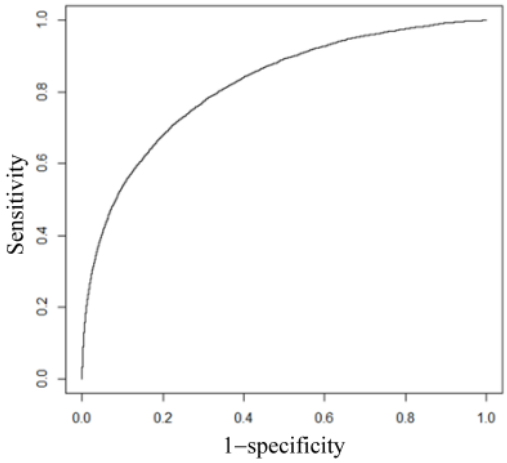


Table 3. The performance measures of our final predictive model when various top percentages of patients having the largest predicted risk were adopted as the cutoff point for doing binary classification.

Top percentage of patients having the largest predicted risk (%)	Accuracy (N=204,744), n (%)	Sensitivity (N=4353), n (%)	Specificity (N=200,391), n (%)	PPV[a]		NPV[b]	
				n (%)	N	n (%)	N
1	199,732 (97.55)	694 (15.94)	199,038 (99.32)	694 (33.90)	2047	199,038 (98.19)	202,697
2	198,349 (96.88)	1026 (23.57)	197,323 (98.47)	1026 (25.06)	4094	197,323 (98.34)	200,650
3	196,831 (96.14)	1291 (29.66)	195,540 (97.58)	1291 (21.02)	6142	195,540 (98.46)	198,602
4	195,186 (95.33)	1492 (34.28)	193,694 (96.66)	1492 (18.22)	8189	193,694 (98.54)	196,555
5	193,472 (94.49)	1659 (38.11)	191,813 (95.72)	1659 (16.21)	10,237	191,813 (98.62)	194,507
6	191,717 (93.64)	1805 (41.47)	189,912 (94.77)	1805 (14.69)	12,284	189,912 (98.68)	192,460
7	189,919 (92.76)	1930 (44.34)	187,989 (93.81)	1930 (13.47)	14,332	187,989 (98.73)	190,412
8	188,124 (91.88)	2056 (47.23)	186,068 (92.85)	2056 (12.55)	16,379	186,068 (98.78)	188,365
9	186,267 (90.98)	2151 (49.41)	184,116 (91.88)	2151 (11.67)	18,426	184,116 (98.82)	186,318
10	184,435 (90.08)	2259 (51.90)	182,176 (90.91)	2259 (11.03)	20,474	182,176 (98.86)	184,270
15	174,902 (85.42)	2611 (59.98)	172,291 (85.98)	2611 (8.50)	30,711	172,291 (99.00)	174,033
20	165,253 (80.71)	2905 (66.74)	162,348 (81.02)	2905 (7.09)	40,948	162,348 (99.12)	163,796
25	155,491 (75.94)	3143 (72.20)	152,348 (76.03)	3143 (6.14)	51,186	152,348 (99.21)	153,558

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

Table 4. The error matrix of our final predictive model when the top 10.00% (20,474/204,744) of patients having the largest predicted risk were adopted as the cutoff point for doing binary classification.

Outcome class	Asthma-related hospital encounters in the following year	No asthma-related hospital encounter in the following year
Projected asthma-related hospital encounters in the following year	2259	18,215
Projected no asthma-related hospital encounter in the following year	2094	182,176

Performance Measures of the Simplified Intermountain Healthcare Model

When applying our simplified Intermountain Healthcare model trained on the Intermountain Healthcare data set [23] to the KPSC test set without retraining the model on the KPSC training set, the model gained an AUC of 0.751 (95% CI 0.742-0.759). When the top 10.00% (20,474/204,744) of patients having the largest predicted risk were adopted as the cutoff point for doing binary classification, the model achieved an accuracy of 89.64% (183,531/204,744; 95% CI 89.51-89.77), a sensitivity of 41.51% (1807/4353; 95% CI 40.14-42.97), a specificity of 90.68% (181,724/200,391; 95% CI 90.55-90.81), a PPV of 8.83% (1807/20,474; 95% CI 8.44-9.23), and an NPV of 98.62% (181,724/184,270; 95% CI 98.57-98.67).

After using the KPSC training set to retrain our simplified Intermountain Healthcare model [23], the model gained on the KPSC test set an AUC of 0.779 (95% CI 0.772-0.787). When the top 10.00% (20,474/204,744) of patients having the largest predicted risk were adopted as the cutoff point for doing binary

classification, the model achieved an accuracy of 89.85% (183,953/204,744; 95% CI 89.71-89.97), a sensitivity of 46.36% (2018/4353; 95% CI 44.89-47.84), a specificity of 90.79% (181,935/200,391; 95% CI 90.65-90.91), a PPV of 9.86% (2018/20,474; 95% CI 9.45-10.25), and an NPV of 98.73% (181,935/184,270; 95% CI 98.68-98.78).

Discussion

Principal Findings

We used KPSC data to develop a model to forecast asthma-related hospital encounters in the following 12-month period in patients with asthma. Table 5 shows that, compared with the models formerly built by others [3,10-22], our final KPSC model gained a higher AUC, that is, our modeling guideline of checking extensive candidate features to boost model accuracy exhibited acceptable generalizability to KPSC. After further enhancement to automatically explain its predictions [40,41] and to raise its accuracy, our model could be used to direct asthma care management to help improve patient outcomes and reduce health care costs.

Asthma affects adults and children differently. Our final model gained a lower AUC in children than in adults. Additional work is required to understand the difference and to boost the prediction accuracy in children.

We examined 337 basic and extended candidate features. Approximately 65.6% (221/337) of these were used in our final model. Many of the unused features were correlated with the outcome variable but provided no additional predictive power on the KPSC data set beyond those used in our final model.

In Multimedia Appendix 1, the 8 most important features and several others within the top 25 features reflect the loss of asthma control. This loss of asthma control could be because of the severity of the patient's asthma. It could also relate to management practices, treatment nonadherence, or socioeconomic factors for which we had no data.

When using our simplified Intermountain Healthcare model [23] without retraining it on the KPSC training set, the model achieved an AUC of 0.751 on the KPSC test set. Despite being 0.069 lower than our final KPSC model's AUC, this AUC is higher than the AUCs of many previous models for predicting hospitalization and ED visits in patients with asthma (Table 5). Therefore, we regard our simplified Intermountain Healthcare model to have acceptable generalizability to KPSC.

Table 5. Our final Kaiser Permanente Southern California model in comparison with several previous models for forecasting hospitalizations and emergency department visits in patients with asthma.

Model	Prediction target	Number of features the model used	Number of data instances	Classification algorithm	The undesirable outcome's prevalence rate in the whole data set (%)	AUC[a]	Sensitivity (%)	Specificity (%)	PPV[b] (%)	NPV[c] (%)
Our final KPSC[d] model	Asthma-related hospital encounters	221	987,506	XGBoost[e]	23,278 (2.36)	0.820	2259 (51.90)	182,176 (90.91)	2259 (11.03)	182,176 (98.86)
Our Intermountain Healthcare model [23]	Asthma-related hospital encounters	142	334,564	XGBoost	12,144 (3.63)	0.859	436 (53.69)	16,955 (91.93)	436 (22.65)	16,955 (97.83)
Miller et al [15]	Asthma-related hospital encounters	17	2821	Logistic regression	8.5	0.81	—[f]	—	—	—
Loymans et al [10]	Asthma exacerbation	7	611	Logistic regression	13	0.8	—	—	—	—
Lieu et al [3]	Asthma-related hospitalization	7	16,520	Proportional hazards regression	1.8	0.79	—	—	—	—
Schatz et al [11]	Asthma-related hospitalization in children	5	4197	Logistic regression	1.4	0.781	43.9	89.8	5.6	99.1
Yurk et al [17]	Lost day or asthma-related hospital encounters	11	4888	Logistic regression	54	0.78	77	63	82	56
Eisner et al [12]	Asthma-related ED[g] visit	3	2415	Logistic regression	18.3	0.751	—	—	—	—
Forno et al [22]	Severe asthma exacerbation	17	615	Scoring	69.6	0.75	—	—	—	—
Schatz et al [11]	Asthma-related hospitalization in adults	3	6904	Logistic regression	1.2	0.712	44.9	87.0	3.9	99.3
Lieu et al [3]	Asthma-related ED visit	7	16,520	Proportional hazards regression	6.4	0.69	—	—	—	—
Eisner et al [12]	Asthma-related hospitalization	1	2858	Logistic regression	32.8	0.689	—	—	—	—
Sato et al [13]	Severe asthma exacerbation	3	78	Classification and regression tree	21	0.625	—	—	—	—
Schatz et al [20]	Asthma-related hospital encounters	4	14,893	Logistic regression	6.5	0.614	25.4	92.0	22.0	93.2
Lieu et al [19]	Asthma-related hospital encounters	4	7141	Classification and regression tree	6.9	—	49.0	83.6	18.5	—

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

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]KPSC: Kaiser Permanente Southern California.

[e]XGBoost: extreme gradient boosting.

[f]The original paper presenting the model did not report the performance measure.

[g]ED: emergency department.

Comparison With Previous Work

Multiple researchers have built models to forecast ED visits and hospitalizations in patients with asthma [3,10-23]. Table 5 compares our final KPSC model with those models, which encompass all pertinent models covered in the systematic review of Loymans et al [18]. With the exception of our Intermountain Healthcare model [23], every model formerly built by others [3,10-22] gained a lower AUC than our final KPSC model. Instead of being for all patients with asthma, the model by Miller et al [15] targets adults with difficult-to-treat or severe asthma, 8.5% of whom had future asthma-related hospital encounters. The model by Loymans et al [10] predicts asthma exacerbations with a prevalence rate of 13%. These 2 prevalence rates of the undesirable outcome are much higher than that in our KPSC data set. In addition, the target patient population and the prediction target of these 2 models are not comparable with those in our KPSC model. Except for these 2 models, each of the other models formerly built by others had an AUC ≤0.79, which is at least 0.030 lower than that of our KPSC model.

Compared with other models, the model by Yurk et al [17] gained a larger PPV and sensitivity mainly because of the use of a distinct prediction target: hospital encounters or one or more days lost because of missed work or reduced activities for asthma. This prediction target was easier to predict, as it occurred in 54% of the patients with asthma. If the model by Yurk et al [17] were used to predict asthma-related hospital encounters that occurred with approximately 2% of the patients with asthma, we would expect the model to gain a lower sensitivity and PPV.

Excluding the model by Yurk et al [17], all of the other models formerly built by others had a sensitivity ≤49%, which is smaller than what our final KPSC model gained: 51.90% (2259/4353). Sensitivity provides, among all patients with asthma who will have future asthma-related hospital encounters, the proportion of patients that the model pinpoints. As the population of patients with asthma is large, for every 1% increase in the identified proportion of patients with asthma who would have future asthma-related hospital encounters, effective care management could help improve patient outcomes, thereby avoiding up to 7200 more ED visits and 1970 more hospitalizations in the United States annually [1,5-8].

The PPV depends substantially on the prevalence rate of undesirable outcomes [42]. In our KPSC test data set, 2.13% (4353/204,744) of patients with asthma had future asthma-related hospital encounters. When the top 10.00% (20,474/204,744) of patients having the largest predicted risk were adopted as the cutoff point for performing binary classification, the maximum possible PPV that a perfect model could obtain is 21.26% (4353/20,474). Our final KPSC model gained a PPV of 11.03% (2259/20,474), which is 51.90% (2259/4353) of the maximum possible PPV. In comparison, in our Intermountain Healthcare test data set, 4.22% of patients with asthma had future asthma-related hospital encounters [23]. Our Intermountain Healthcare model gained a PPV of 22.65% (436/1925) [23], which is 53.7% (436/812) of the maximum possible PPV that a perfect model could obtain. On a data set in which 6.5% of patients with asthma had future asthma-related hospital encounters, the model by Schatz et al [20] gained a PPV of 22.0%. On a data set in which 6.9% of patients with asthma had future asthma-related hospital encounters, the model by Lieu et al [19] gained a PPV of 18.5%. Except for these PPVs and the PPV of the model by Yurk et al [17], none of the previously reported PPVs was more than 5.6%.

Despite being built using the same modeling guideline, our final KPSC model gained a lower AUC than our Intermountain Healthcare model [23]. This is largely because the percentage of data instances in the test set linking to future asthma-related hospital encounters differs greatly at Intermountain Healthcare and at KPSC: 4.22% (812/19,256) versus 2.13% (4353/204,744), respectively. The rarer the undesirable outcome, the harder it is to accurately predict it.

The top features with an importance value ≥0.01 in our final KPSC model are similar to those in our Intermountain Healthcare model [23]. In both our final KPSC and our Intermountain Healthcare models, many top features involve asthma medications and previous ED visits. When building our Intermountain Healthcare model, we did not consider several basic candidate features. They turned out to be top features in our final KPSC model and impacted the importance values and ranks of the other top features there.

When building our Intermountain Healthcare model, we did not incorporate any extended candidate features. Several such features appeared as top features in our final KPSC model. Their inclusion boosted the model accuracy on our KPSC data set. It is possible that including extended candidate features could also boost the model accuracy on our Intermountain Healthcare data set. This could be explored in future work.

Schatz et al [20] showed that in 2 Southern California cities, 6.5% of patients with asthma at KPSC had asthma-related hospital encounters in 2000. In comparison, 2.08% (4353/208,959) of patients with asthma at KPSC had asthma-related hospital encounters in 2018. This suggests that compared with 2 decades ago, KPSC manages patients with asthma better now.

Considerations About Potential Clinical Use

Although more accurate than those formerly built by others, our final KPSC model still gained a somewhat low PPV of 11.03% (2259/20,474). However, our model could be clinically useful:

1. A PPV of 11.03% (2259/20,474) is acceptable for pinpointing high-risk patients with asthma to apply low-cost preventive interventions. Examples of such interventions include giving the patient a peak flow meter for self-monitoring at home and showing the patient how to use it, instructing the patient on the correct use of an asthma inhaler, asking a nurse to follow up on the patient with extra

phone calls, and training the patient to write a diary on environmental triggers.

2. As explained above, because of the low prevalence rate of the undesirable outcome used in this study, even a perfect model would gain a small PPV. For this outcome, sensitivity matters more than PPV for judging the model's possible clinical impact. Our final KPSC model gained a higher sensitivity than all of the models that were formerly built by others and used a comparable prediction target.
3. To allocate care management resources, health care systems such as the University of Washington Medicine, Kaiser Permanente Northern California [3], and Intermountain Healthcare are using proprietary models whose performance measures are akin to those of the models previously built by others. Our final KPSC model is more accurate than these models.

Our final KPSC model used 221 features. Cutting this number could facilitate the clinical deployment of the model. In this regard, if one could bear a small drop in prediction accuracy, one could adopt the top features having an importance value of, for example, 0.01 or more and remove the others. The importance value of a feature changes across health care systems. Ideally, before deciding which features to keep, one should first compute the importance values of the features on a data set from the intended health care system.

Most of the attributes that we used to compute the features adopted in our final KPSC model, particularly the top features, are routinely collected by electronic medical record systems. For future work, to make it easy for other health care systems to reuse our final KPSC model, we can resort to the Observational Medical Outcomes Partnership (OMOP) common data model [43]. This data model and its linked standardized terminologies [44] standardize administrative and clinical attributes from at least 10 large US health care systems [45,46]. We can extend this data model to include the attributes that are used in our final KPSC model but missed by the original data model. We rewrite our feature construction and model building code based on the extended OMOP common data model and post our code and the related data schema on a public website. After converting its data into our extended OMOP common data model format based on this data schema, a health care system can rerun our code on its data to obtain a simplified version of our final KPSC model tailored to its data. Hopefully, most of the predictive power of our final KPSC model can be retained similar to what this study showed for our Intermountain Healthcare model.

It is difficult to interpret an XGBoost model employing many features globally, as is the case with many other involved machine learning models. As an interesting topic for future work, we plan to use our previously proposed method [40,41] to automatically explain our final KPSC model's predictions for each patient with asthma.

Our final KPSC model was an XGBoost model [32]. When classifying 2 unbalanced classes, XGBoost employs a hyperparameter scale_pos_weight to balance their weights [47]. To maximize the AUC of our KPSC model, our automatic model selection method [37] changed scale_pos_weight from its default value to balance the 2 classes of having future asthma-related hospital encounters or not [48]. As a side effect, this shrank the model's projected probabilities of having future asthma-related hospital encounters to a large extent and made them differ greatly from the actual probabilities [48]. This does not affect the identification of the top few percent of patients with asthma who have the largest projected risk to receive care management or other preventive interventions. We could keep scale_pos_weight at its default value of 1 and not balance the 2 classes. This would avoid the side effect but drop the model's AUC from 0.820 to 0.817 (95% CI 0.810-0.824).

Limitations

This study has 3 limitations, all of which provide interesting areas for future work:

1. In addition to those examined in this study, other features could also help raise model accuracy. Our KPSC data set does not include some potentially relevant features, such as characteristics of the patient's home environment and features computed on the data gathered by monitoring sensors attached to the patient's body. It would be worthwhile to identify new predictive features from various data sources.
2. Our study used only non-deep learning machine learning algorithms and structured data. Using deep learning and including features computed from unstructured clinical notes may further boost model accuracy [41,49].
3. Our study assessed our modeling guideline's generalizability to only one health care system. It would be interesting to evaluate our modeling guideline's generalizability to other health care systems, such as academic health care systems that have different properties from KPSC and Intermountain Healthcare. Compared with nonacademic health care systems, academic health care systems tend to care for sicker and more complex patients [50]. To perform such an evaluation, we are working on obtaining a data set of patients with asthma from the University of Washington Medicine [49].

Conclusions

In its first generalizability assessment, our modeling guideline of examining extensive candidate features to help boost model accuracy exhibited acceptable generalizability to KPSC. Compared with the models formerly built by others, our KPSC model for projecting asthma-related hospital encounters in patients with asthma gained a higher AUC. At present, predictive models are widely used as a core component of a decision support tool to prospectively pinpoint high-risk patients with asthma for preventive care via care management. After further enhancement, our KPSC model could be used to replace the existing predictive models in the decision support tool for better directing asthma care management to help improve patient outcomes and reduce health care costs.

Acknowledgments

The authors would like to thank Lee J Barton, Don McCarthy, Xia X Li, and Michael D Johnson for useful discussions and helping to retrieve the KPSC data set. GL, CN, MS, RZ, ER, and CK were partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL142503. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

GL was mainly responsible for this study. He conceptualized and designed the study, performed the literature review and data analysis, and wrote the paper. CK, CN, WC, MS, ER, and RZ provided feedback on various medical issues, contributed to conceptualizing the presentation, and revised the paper. CK and CN took part in retrieving the KPSC data set and interpreting its detected peculiarities.

Conflicts of Interest

RZ reports grants from Aerocrine, grants and personal fees from Genentech, grants and personal fees from MedImmune of AstraZeneca, grants and personal fees from Merck, personal fees from Novartis, personal fees from Regeneron Pharmaceuticals, grants and personal fees from GlaxoSmithKline, grants from ALK Pharma, and grants from TEVA Pharmaceutical Industries Ltd outside this study.

Multimedia Appendix 1
The basic candidate features, the clinical and demographic characteristics of our patient cohort, and the features employed in our final predictive model and their importance values.
[PDF File (Adobe PDF File), 171 KB - medinform_v8i11e22689_app1.pdf]

References

1. Moorman JE, Akinbami LJ, Bailey CM, Zahran HS, King ME, Johnson CA, et al. National surveillance of asthma: United States, 2001-2010. Vital Health Stat 3 2012 Nov(35):1-58 [FREE Full text] [Medline: 24252609]
2. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc 2018 Mar;15(3):348-356. [doi: 10.1513/AnnalsATS.201703-259OC] [Medline: 29323930]
3. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998 Apr;157(4 Pt 1):1173-1180. [doi: 10.1164/ajrccm.157.4.9708124] [Medline: 9563736]
4. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:W4-427. [doi: 10.1377/hlthaff.w4.427] [Medline: 15451964]
5. Caloyeras JP, Liu H, Exum E, Broderick M, Mattke S. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. Health Aff (Millwood) 2014 Jan;33(1):124-131. [doi: 10.1377/hlthaff.2013.0625] [Medline: 24395944]
6. Greineder DK, Loane KC, Parks P. A randomized controlled trial of a pediatric asthma outreach program. J Allergy Clin Immunol 1999 Mar;103(3 Pt 1):436-440. [doi: 10.1016/s0091-6749(99)70468-9] [Medline: 10069877]
7. Kelly CS, Morrow AL, Shults J, Nakas N, Strope GL, Adelman RD. Outcomes evaluation of a comprehensive intervention program for asthmatic children enrolled in medicaid. Pediatrics 2000 May;105(5):1029-1035. [doi: 10.1542/peds.105.5.1029] [Medline: 10790458]
8. Axelrod RC, Zimbro KS, Chetney RR, Sabol J, Ainsworth VJ. A disease management program utilizing life coaches for children with asthma. J Clin Outcomes Manag 2001;8(6):38-42 [FREE Full text]
9. Axelrod RC, Vogel D. Predictive modeling in health plans. Dis Manag Health Outcomes 2003;11(12):779-787. [doi: 10.2165/00115677-200311120-00003]
10. Loymans RJ, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJ, Schermer TRJ, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016 Sep;71(9):838-846. [doi: 10.1136/thoraxjnl-2015-208138] [Medline: 27044486]
11. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003 Aug;9(8):538-547 [FREE Full text] [Medline: 12921231]
12. Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest 2012 Jan;141(1):58-65. [doi: 10.1378/chest.11-0020] [Medline: 21885725]
13. Sato R, Tomita K, Sano H, Ichihashi H, Yamagata S, Sano A, et al. The strategy for predicting future exacerbation of asthma using a combination of the asthma control test and lung function test. J Asthma 2009 Sep;46(7):677-682. [doi: 10.1080/02770900902972160] [Medline: 19728204]

14. Osborne ML, Pedula KL, O'Hollaren M, Ettinger KM, Stibolt T, Buist AS, et al. Assessing future need for acute care in adult asthmatics: the profile of asthma risk study: a prospective health maintenance organization-based study. Chest 2007 Oct;132(4):1151-1161. [doi: 10.1378/chest.05-3084] [Medline: 17573515]
15. Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, TENOR Study Group. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006 Dec;28(6):1145-1155. [doi: 10.1183/09031936.06.00145105] [Medline: 16870656]
16. Peters D, Chen C, Markson LE, Allen-Ramey FC, Vollmer WM. Using an asthma control questionnaire and administrative data to predict health-care utilization. Chest 2006 Apr;129(4):918-924. [doi: 10.1378/chest.129.4.918] [Medline: 16608939]
17. Yurk RA, Diette GB, Skinner EA, Dominici F, Clark RD, Steinwachs DM, et al. Predicting patient-reported asthma outcomes for adults in managed care. Am J Manag Care 2004 May;10(5):321-328 [FREE Full text] [Medline: 15152702]
18. Loymans RJ, Debray TP, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TR, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-52.e15. [doi: 10.1016/j.jaip.2018.02.004] [Medline: 29454163]
19. Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty of half full? J Asthma 1999 Jun;36(4):359-370. [doi: 10.3109/02770909909068229] [Medline: 10386500]
20. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004 Jan;10(1):25-32 [FREE Full text] [Medline: 14738184]
21. Grana J, Preston S, McDermott PD, Hanchak NA. The use of administrative data to risk-stratify asthmatic patients. Am J Med Qual 1997;12(2):113-119. [doi: 10.1177/0885713X9701200205] [Medline: 9161058]
22. Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, et al. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010 Nov;138(5):1156-1165 [FREE Full text] [doi: 10.1378/chest.09-2426] [Medline: 20472862]
23. Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. JMIR Med Inform 2020 Jan 21;8(1):e16080 [FREE Full text] [doi: 10.2196/16080] [Medline: 31961332]
24. Mayfield J, McNamee P, Piatko CD. Named Entity Recognition Using Hundreds of Thousands of Features. In: Proceedings of the Seventh Conference on Natural Language Learning. 2003 Presented at: CoNLL'03; May 31-June 1, 2003; Edmonton, Canada. [doi: 10.3115/1119176.1119205]
25. Cao Y, Yu H, Abbott NL, Zavala VM. Machine learning algorithms for liquid crystal-based sensors. ACS Sens 2018 Nov 26;3(11):2237-2245. [doi: 10.1021/acssensors.8b00100] [Medline: 30289249]
26. Zhai Y, Ong Y, Tsang IW. The emerging 'big dimensionality'. IEEE Comput Intell Mag 2014 Aug;9(3):14-26. [doi: 10.1109/mci.2014.2326099]
27. Hansson K, Yella S, Dougherty M, Fleyeh H. Machine learning algorithms in heavy process manufacturing. Am J Intell Syst 2016;6(1):1-13. [doi: 10.5923/j.ajis.20160601.01]
28. Koebnick C, Langer-Gould AM, Gould MK, Chao CR, Iyer RL, Smith N, et al. Sociodemographic characteristics of members of a large, integrated health care system: comparison with US Census Bureau data. Perm J 2012;16(3):37-41 [FREE Full text] [doi: 10.7812/tpp/12-031] [Medline: 23012597]
29. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Med Care 2012 Jul(50 Suppl):S30-S35 [FREE Full text] [doi: 10.1097/MLR.0b013e318259c011] [Medline: 22692256]
30. Wakefield DB, Cloutier MM. Modifications to HEDIS and CSTE algorithms improve case recognition of pediatric asthma. Pediatr Pulmonol 2006 Oct;41(10):962-971. [doi: 10.1002/ppul.20476] [Medline: 16871628]
31. Andrews AL, Simpson AN, Basco WT, Teufel RJ. Asthma medication ratio predicts emergency department visits and hospitalizations in children with asthma. Medicare Medicaid Res Rev 2013;3(4) [FREE Full text] [doi: 10.5600/mmrr.003.04.a05] [Medline: 24834366]
32. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'16; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
33. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. New York, USA: Springer; 2016.
34. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Second Edition. New York, USA: Springer; 2019.
35. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition. Burlington, MA: Morgan Kaufmann; 2016.
36. XGBoost JVM package. 2020. URL: https://xgboost.readthedocs.io/en/latest/jvm/index.html [accessed 2020-10-31]

37. Zeng X, Luo G. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Health Inf Sci Syst 2017 Dec;5(1):2 [FREE Full text] [doi: 10.1007/s13755-017-0023-z] [Medline: 29038732]
38. Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013 Presented at: KDD'13; August 11-14, 2013; Chicago, IL. [doi: 10.1145/2487575.2487629]
39. Agresti A. Categorical Data Analysis. Third Edition. Hoboken, NJ: Wiley; 2012.
40. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst 2016;4:2 [FREE Full text] [doi: 10.1186/s13755-016-0015-4] [Medline: 26958341]
41. Luo G. A roadmap for semi-automatically extracting predictive and clinically meaningful temporal features from medical data for predictive modeling. Glob Transit 2019;1:61-82 [FREE Full text] [doi: 10.1016/j.glt.2018.11.001] [Medline: 31032483]
42. Ranganathan P, Aggarwal R. Common pitfalls in statistical analysis: understanding the properties of diagnostic tests – part 1. Perspect Clin Res 2018;9(1):40-43. [doi: 10.4103/picr.picr_170_17]
43. Observational Medical Outcomes Partnership (OMOP) Common Data Model. 2020. URL: http://omop.org/CDM [accessed 2020-10-28]
44. Vocabularies. Observational Medical Outcomes Partnership (OMOP). 2020. URL: http://omop.org/Vocabularies [accessed 2020-10-28]
45. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
46. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
47. Parameters. XGBoost. 2020. URL: https://xgboost.readthedocs.io/en/latest/parameter.html [accessed 2020-10-28]
48. Notes on Parameter Tuning. XGBoost. 2020. URL: https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html [accessed 2020-10-28]
49. Luo G, Stone BL, Koebnick C, He S, Au DH, Sheng X, et al. Using temporal features to provide data-driven clinical early warnings for chronic obstructive pulmonary disease and asthma care management: protocol for a secondary analysis. JMIR Res Protoc 2019 Jun 6;8(6):e13783 [FREE Full text] [doi: 10.2196/13783] [Medline: 31199308]
50. Liu LL, Forgione DA, Younis MZ. A comparative analysis of the CVP structure of nonprofit teaching and for-profit non-teaching hospitals. J Health Care Finance 2012;39(1):12-38 [FREE Full text]

Abbreviations

AUC: area under the receiver operating characteristic curve
ED: emergency department
FN: false negative
FP: false positive
ICD: International Classification of Diseases
KPSC: Kaiser Permanente Southern California
NPV: negative predictive value
OMOP: Observational Medical Outcomes Partnership
PCP: primary care provider
PPV: positive predictive value
ROC: receiver operating characteristic
TN: true negative
TP: true positive
WEKA: Waikato Environment for Knowledge Analysis
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 29.07.20; peer-reviewed by T Agresta, C Fuller; comments to author 06.09.20; revised version received 15.09.20; accepted 18.10.20; published 09.11.20.

Please cite as:
Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Rozema E, Koebnick C
Developing a Predictive Model for Asthma-Related Hospital Encounters in Patients With Asthma in a Large, Integrated Health Care System: Secondary Analysis
JMIR Med Inform 2020;8(11):e22689
URL: http://medinform.jmir.org/2020/11/e22689/
doi:10.2196/22689
PMID:33164906

Review

Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis

Herdiantri Sufriyana[1,2], MSc, MD; Atina Husnayain[1,3], MPH; Ya-Lin Chen[1,4], PharmD; Chao-Yang Kuo[1], MSc; Onkar Singh[5,6], MSc; Tso-Yang Yeh[7]; Yu-Wei Wu[1,8], PhD; Emily Chia-Yu Su[1,8], PhD

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

[2]Department of Medical Physiology, College of Medicine, University of Nahdlatul Ulama Surabaya, Surabaya, Indonesia

[3]Department of Biostatistics, Epidemiology, and Population Health, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia

[4]School of Pharmacy, College of Pharmacy, Taipei Medical University, Taipei, Taiwan

[5]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, Taiwan

[6]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan

[7]School of Dentistry, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan

[8]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan

Corresponding Author:
Emily Chia-Yu Su, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
250 Wu-Xing Street
Taipei, 11031
Taiwan
Phone: 886 2 663 82736 ext 1515
Email: emilysu@tmu.edu.tw

Abstract

Background: Predictions in pregnancy care are complex because of interactions among multiple factors. Hence, pregnancy outcomes are not easily predicted by a single predictor using only one algorithm or modeling method.

Objective: This study aims to review and compare the predictive performances between logistic regression (LR) and other machine learning algorithms for developing or validating a multivariable prognostic prediction model for pregnancy care to inform clinicians' decision making.

Methods: Research articles from MEDLINE, Scopus, Web of Science, and Google Scholar were reviewed following several guidelines for a prognostic prediction study, including a risk of bias (ROB) assessment. We report the results based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Studies were primarily framed as PICOTS (population, index, comparator, outcomes, timing, and setting): Population: men or women in procreative management, pregnant women, and fetuses or newborns; Index: multivariable prognostic prediction models using non-LR algorithms for risk classification to inform clinicians' decision making; Comparator: the models applying an LR; Outcomes: pregnancy-related outcomes of procreation or pregnancy outcomes for pregnant women and fetuses or newborns; Timing: pre-, inter-, and peripregnancy periods (predictors), at the pregnancy, delivery, and either puerperal or neonatal period (outcome), and either short- or long-term prognoses (time interval); and Setting: primary care or hospital. The results were synthesized by reporting study characteristics and ROBs and by random effects modeling of the difference of the logit area under the receiver operating characteristic curve of each non-LR model compared with the LR model for the same pregnancy outcomes. We also reported between-study heterogeneity by using τ^2 and I^2.

Results: Of the 2093 records, we included 142 studies for the systematic review and 62 studies for a meta-analysis. Most prediction models used LR (92/142, 64.8%) and artificial neural networks (20/142, 14.1%) among non-LR algorithms. Only 16.9% (24/142) of studies had a low ROB. A total of 2 non-LR algorithms from low ROB studies significantly outperformed

LR. The first algorithm was a random forest for preterm delivery (logit AUROC 2.51, 95% CI 1.49-3.53; I^2=86%; τ^2=0.77) and pre-eclampsia (logit AUROC 1.2, 95% CI 0.72-1.67; I^2=75%; τ^2=0.09). The second algorithm was gradient boosting for cesarean section (logit AUROC 2.26, 95% CI 1.39-3.13; I^2=75%; τ^2=0.43) and gestational diabetes (logit AUROC 1.03, 95% CI 0.69-1.37; I^2=83%; τ^2=0.07).

Conclusions: Prediction models with the best performances across studies were not necessarily those that used LR but also used random forest and gradient boosting that also performed well. We recommend a reanalysis of existing LR models for several pregnancy outcomes by comparing them with those algorithms that apply standard guidelines.

Trial Registration: PROSPERO (International Prospective Register of Systematic Reviews) CRD42019136106; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=136106

(JMIR Med Inform 2020;8(11):e16503) doi:10.2196/16503

KEYWORDS

machine learning; pregnancy complications; prognosis; clinical prediction rule; meta-analysis; systematic review

Introduction

Background

Pregnancy is a common health condition that requires long-term rigorous care to anticipate adverse outcomes. Most pregnancy outcomes are identified after delivery; however, these are results of interactions among multiple factors occurring for many weeks beforehand. The number of factors and their interactions along with the time intervals make predictions of pregnancy outcomes very complicated. Multiple or multivariable logistic regression (LR) is widely used to deal with similar multifactorial problems in health outcome research [1]. Applied to medicine, statistics, and machine learning (computer science), this algorithm fits multiple parameters in a prediction model by assuming that predictors are linearly and additively related to an outcome [2]. Nevertheless, nonlinear problems commonly occur in human physiology because of complex interactions, such that a linear model might not be capable of adequately predicting outcomes [3]. With the growth of machine learning applications in health care, applying other algorithms may scale up the solution space for accurate predictions of pregnancy outcomes long before giving birth.

Despite improvements in maternal and neonatal mortality, conditions still differ between developing and developed countries or regions [4]. The most common causes of maternal deaths are hemorrhage, hypertension, and sepsis [5], whereas the causes of neonatal deaths are mostly due to prematurity, birth asphyxia, and infections [6]. Postpartum hemorrhage and sepsis are further compounded by multiple causes and risk factors [7,8], and hypertension in pregnancy or prematurity is associated with multiple mechanisms [9,10]. The aforementioned diseases and complications cannot be very easily predicted by a single epidemiological predictor, a single measure by a medical device, or a single biomarker. Furthermore, interactions among multiple predictors also might not be captured by a single machine learning algorithm including LR. Therefore, a prediction study may need to compare multiple machine learning algorithms to develop a prognostic prediction model that uses multiple predictors.

Machine learning algorithms have long been applied for clinical prediction purposes. A support vector machine demonstrated a summary of receiver operating characteristics (ROCs) of >90% for breast cancer prognostic prediction [11]. To predict therapeutic outcomes in depression, the pooled estimated accuracy of machine learning algorithms was 0.82 (95% CI 0.77-0.87) [12]. However, the difference in the logit area under the ROC curve (AUROC) was 0.00 (95% CI −0.18 to 0.18) between LR and machine learning in studies with a low risk of bias (ROB) [13]. A similar conclusion was found for predicting intracerebral hemorrhage (P=.49) outlined in a systematic review [14]. These previous results imply that (1) machine learning algorithms may or may not perform better than traditional modeling by LR and (2) applying only a single algorithm may cause an investigator to lose the chance to obtain a model with optimal predictive performance using the same predictors. Meanwhile, a unique interaction should exist between a set of predictors and a pregnancy outcome. A particular predictive algorithm may work best to capture this predictor-outcome interaction. Prediction tasks are even more challenging in pregnancy care because they demand more prognostic instead of diagnostic predictions. Yet, unlike the common nature of other long-term conditions in health care (eg, diabetes mellitus), the onset, time to event, and target population in pregnancy care are rather apparent. However, unpredictable events leading to disabilities and death in a population such as pregnant women or newborns are also not easily accepted as in other populations (eg, patients with cancer and older adults). Thus, clinicians should apply several prediction models with satisfactory predictive performances throughout the pregnancy period. Clinicians and investigators would benefit from knowing whether an LR or other algorithms have a better chance of achieving satisfactory predictive performances for a particular pregnancy outcome. However, no previous systematic review in pregnancy care has reviewed multiple machine learning algorithms and compared their predictive performances, including LR, to predict pregnancy outcomes.

This review will allow investigators and clinicians in pregnancy care to consider the development or application of prediction models throughout the pregnancy period. This review demonstrates which algorithms have shown robust predictive performances for a particular pregnancy outcome using a similar set of predictors. Investigators in pregnancy care may also consider whether a reanalysis by another predictive algorithm

is needed by using existing data previously analyzed by an algorithm including LR. Beyond the algorithm issue, the development of machine learning models also requires an adequate methodology and interpretable results [15]. Biased conclusions should be avoided when describing machine learning predictive performances [11,16]. Standard guidelines are important when investigating and reviewing machine learning applications in clinical prediction modeling [15,17].

Objectives

By applying the standard guidelines, we aim to review machine learning models and compare their predictive performances between LRs and other machine learning algorithms. In this review, we focus on machine learning models either developed or validated for making prognostic predictions in pregnancy care intended to inform clinicians' decision making.

Methods

Protocol and Registration

We reported this study based on PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [18] and conducted the review based on several guidelines related to prediction studies. The review objective was defined according to a standard of key items [19]. Our eligibility criteria were composed of items elaborated with 2 guidelines for developing and reporting a prediction model and a guideline for assessing the applicability. These included transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [20] and another that focuses on machine learning modeling in biomedical research (hereafter referred to as guidelines for developing and reporting machine learning predictive models in biomedical research [MLP-BIOM]) [15]. Applicability was assessed using assessments that were a part of the *prediction model risk of bias assessment tool* (PROBAST) [17,21]. Data were extracted based on the checklist for critical appraisal and data extraction for systematic reviews of prediction modeling studies (CHARMS), which also describes items for the review objective. Our review protocol was registered with PROSPERO (CRD42019136106).

Eligibility Criteria

Before defining the eligibility criteria, we decided to view the LR as one of many algorithms in the machine learning field with respect to its use in statistics and data science. A prediction model development consisted of several elements: predictor selection, parameter fitting, and hyperparameter optimization [2]. In this review, the term *prediction model* refers to all those elements, whereas the term *prediction algorithm* refers to a parameter-fitting method. Using the same set of predictors, we would expect different predictive performances if the parameters of a model are fitted using different algorithms. A prediction algorithm in machine learning is a way for the computer to learn from data by fitting the parameters with respect to predicting a class measured by hyperparameters from the human user [22]. Several optimization algorithms have been developed to reduce the human role in determining these hyperparameters, such as sequential search, random search, and Bayesian optimization [23]. However, that is beyond the scope of this review.

By focusing on prediction algorithms, we defined eligibility criteria to screen studies by the title, abstract, and full text. We also assessed the applicability by examining the full text. These were the candidates we selected for the qualitative analysis. Key items of population, index, comparator, outcomes, timing, setting (PICOTS) [19] and additional items [15,20] composed the eligibility criteria. The first item of these criteria was a review question framed using PICOTS. The key items consisted of the following:

1. Population: men or women in procreative management, pregnant women, and fetuses or newborns.
2. Index: multivariable prognostic prediction models applying non-LR algorithms for risk classification tasks intended to inform clinicians' decision making.
3. Comparator: multivariable prognostic predictions applying an LR algorithm, excluding a scoring system in which the parameters determined by humans instead of using LR, for risk classification tasks intended to inform clinicians' decision making.
4. Outcomes: pregnancy-related outcomes of procreative management or pregnancy outcomes for pregnant women or fetuses or newborns.
5. Timing: with predictors being measured at the pre-, inter-, and peripregnancy periods and outcomes being assessed at the pregnancy, delivery, and either puerperal or neonatal period, short- and long-term prognoses were applied.
6. Setting: primary care or hospital.

Additional items were the availability of several reporting components as required by TRIPOD and MLP-BIOM. These components included (1) data sources, (2) outcomes, (3) evaluation metrics, (4) predictors, (5) descriptive statistics, (6) event sample sizes, (7) modeling methods or algorithms, and (8) model validation.

After briefly screening studies by eligibility criteria, we conducted an applicability assessment by thoroughly examining the full texts. Using PROBAST guidelines, we assessed the applicability according to the review question framed by PICOTS. Low, high, or unclear criteria were determined for applicable, not applicable, or unclear applicability, respectively. The assessment covered 3 domains of participants, predictors, and outcomes. Only those fulfilling *low* criteria were selected for the qualitative analysis.

For the quantitative analysis, studies had to report the AUROC. Studies were selected from those applicable for the qualitative analysis. If there were at least three LR models and a non-LR model from any studies for an outcome, all studies with that outcome were included in the meta-analysis. This was determined based on the requirement of a minimum number of data points to calculate the variance as part of the meta-analytical procedure. If studies did not report the AUROC, we estimated the sensitivity and specificity using the trapezoidal rule (see *Summary Measures* and *Synthesis of Results* sections).

Information Sources

We searched the MEDLINE, Scopus, Web of Science, and Google Scholar databases up to May 2020. There was no limit on the publication period. However, considering the limitations

of the search interface in Google Scholar, we only retrieved results from the last year with keywords in the abstract or the entire period with those keywords in the title. We also limited the publication period to the last 10 years for search results by keywords including "logistic regression multivariable prediction." This was because we estimated that there would be enormous amounts of studies applying LR because we applied a broad range of outcomes in this study. In contrast, we might lack studies using other machine learning models, although the outcomes were broad.

Search

The initial search filter was limited to the title, abstract, keywords, or Medical Subject Heading (MeSH; MEDLINE only) using "machine learning" AND pregnancy. We also used "machine learning AND ([pregnancy outcome from initial search] NOT pregnancy)." Keywords for pregnancy outcomes were used based on MeSH to generalize a variety of terms for pregnancy outcomes from selected studies. If the MeSH term contained "pregnancy," then we used the alternative entry terms in the webpage recorded for this MeSH term. If all entry terms also contained "pregnancy," then we used the term without negating "pregnancy." In addition, we also substituted the "machine learning" part with one of the keywords consisting of "decision tree," "artificial neural network," "support vector machine," "random forest," "artificial intelligence," "deep learning," and "logistic regression multivariable prediction." All keywords are described in Multimedia Appendix 1. These search terms were applied to all databases.

Study Selection

Duplicate records from multiple databases were removed. We refined the search results in the title or abstract using EndNote X8 (Clarivate Analytics) by "(supervised NOT unsupervised) OR prediction OR classification." Records were screened by HS and AH, and the results were assessed by HS, AH, YC, CK, OS, TY, and YW. Disagreements were resolved by discussion with the last author (ES). Study selection was conducted in brief and thorough assessments. These brief assessments were intended to select studies by checking eligibility criteria from TRIPOD and MLP-BIOM in the title, abstract, and briefly in the full-text article. A thorough assessment of the applicability from PROBAST was conducted later before the ROB assessment.

Data Collection Process

We extracted data based on the CHARMS checklist, which includes (1) outcomes, (2) study design, (3) data sources, (4) data source design, (5) setting, (6) type of study, and (7) modeling methods or algorithms, and (8) predictive performance. Outcomes were pooled as distinct MeSH terms. Study and data source designs were classified into prospective, retrospective, nested case-control, case-control, and cross-sectional. We defined the type of study based on the model validation, which might be development, validation, or both. Eligible studies were described as developing prediction models by applying LR, non-LR, or both algorithms. Predictive performances were only taken from studies that were eligible for the meta-analysis (see *Eligibility Criteria* section). If there were multiple models developed within a study using the same algorithm, we retrieved the AUROC from the best performing one among the models. If both LR and non-LR algorithms were applied in a study, we selected the predictive performances of the best models applying either the LR or non-LR algorithm. Model performances derived from external validation were preferred if available.

ROB Within and Across Studies

We used PROBAST to assess the ROB [17,21]. The ROB in individual studies was assessed as low, high, or unclear in 4 domains of participants, predictors, outcomes, and analyses. In addition, 20 signaling questions were answered for each study in a transparent and accountable form. Across studies, we described the proportion of low, unclear, or high ROBs. ROBs were compared for each domain. We also summarized the answers for each signaling question.

Summary Measures

We compared AUROCs from studies that reported this metric. Logit transformation was applied to the AUROCs. We computed logit AUROC differences between each non-LR and LR algorithm across studies. Summary measures from any eligible studies with all, low, or high ROB were pooled by random effects modeling, as previously described [24]. Assuming that selected studies were random samples from a larger population, we chose a random effects model that attempted to generalize findings beyond the included studies using that assumption [25]. Despite this, we did not conduct random effects modeling for all selected studies considering the broad range of target populations, outcomes, and algorithms. Meanwhile, we conducted this review within a narrower field compared with a previous systematic review of machine learning in medicine [13]. Therefore, we only applied random effects modeling to the predictive performances of selected studies using a particular pregnancy outcome. These studies consisted of a minimum number of non-LR and 3 LR models from any studies. This minimum number was considered to obtain a minimum number of data points of logit AUROCs to compute the interval estimates in a random effects model. We depicted the AUROCs using forest plots; thus, one can see which prediction algorithm may have a better chance of obtaining optimal predictive performance for a particular pregnancy outcome.

Pooled estimates of pairwise differences in logit AUROCs were described by points and the 95% CI [26]. A positive difference in logit AUROCs means that the non-LR algorithm had a higher logit AUROC than that of the LR algorithm. The difference was significant if 0 was not included within the 95% CI. The number of pairwise comparisons (k) for each random effects model was reported. We also reported variance across studies (τ^2) and I^2 as absolute and relative values of between-study heterogeneity, respectively.

If a study did not report the AUROC, we estimated this metric based on sensitivity and specificity. As a specificity of 0% means a sensitivity of 100% and *vice versa*, the AUROC could be estimated from the reported sensitivity and specificity using a common rule to calculate the area of the trapezoid (Equation 1). Before we subtracted the AUROC of a non-LR algorithm

from that of an LR algorithm, we applied a logit transformation (Equation 2).

AUROC = 0.5 × (1 − specificity) × sensitivity + specificity × sensitivity + 0.5 × (1 − sensitivity) × specificity **(1)**

Logit(AUROC) = log (AUROC / (1 – AUROC)) **(2)**

We used RStudio 1.2 (RStudio) with R 3.6.1 and an additional package, metafor 2.4.0, for random effects modeling. We applied the restricted maximum likelihood estimator method [27]. These are common tools and recommended modeling methods for meta-analyses [28].

Synthesis of Results

We described the characteristics of the studies consisting of population, study design, timing, and setting. This was described as the number of algorithms used for prediction modeling. The algorithms were categorized into LR, non-LR, or both algorithms. We also show the proportion of each characteristic compared with all characteristics within the same algorithm category.

ROBs within studies were described for the number of low, high, or unclear ROB studies. This was reported for overall assessment results and by domain in studies that used LR, non-LR, or both algorithms. ROBs across studies were described for the proportion of studies in which the answer to each signaling question led to low, high, or unclear ROB studies. We intended to show what makes most studies considered to have high ROBs.

Meta-analytical results were described by a forest plot faceted by outcome. Each facet showed comparisons of differences in logit AUROCs for each random effects model of non-LR versus LR algorithms. This demonstrated which algorithms tended to outperform LR for each pregnancy outcome. Comparisons that included non-LR high ROB studies were color coded. The best predictive performance for each outcome was reported. Between-study heterogeneity for each random effects model was also reported.

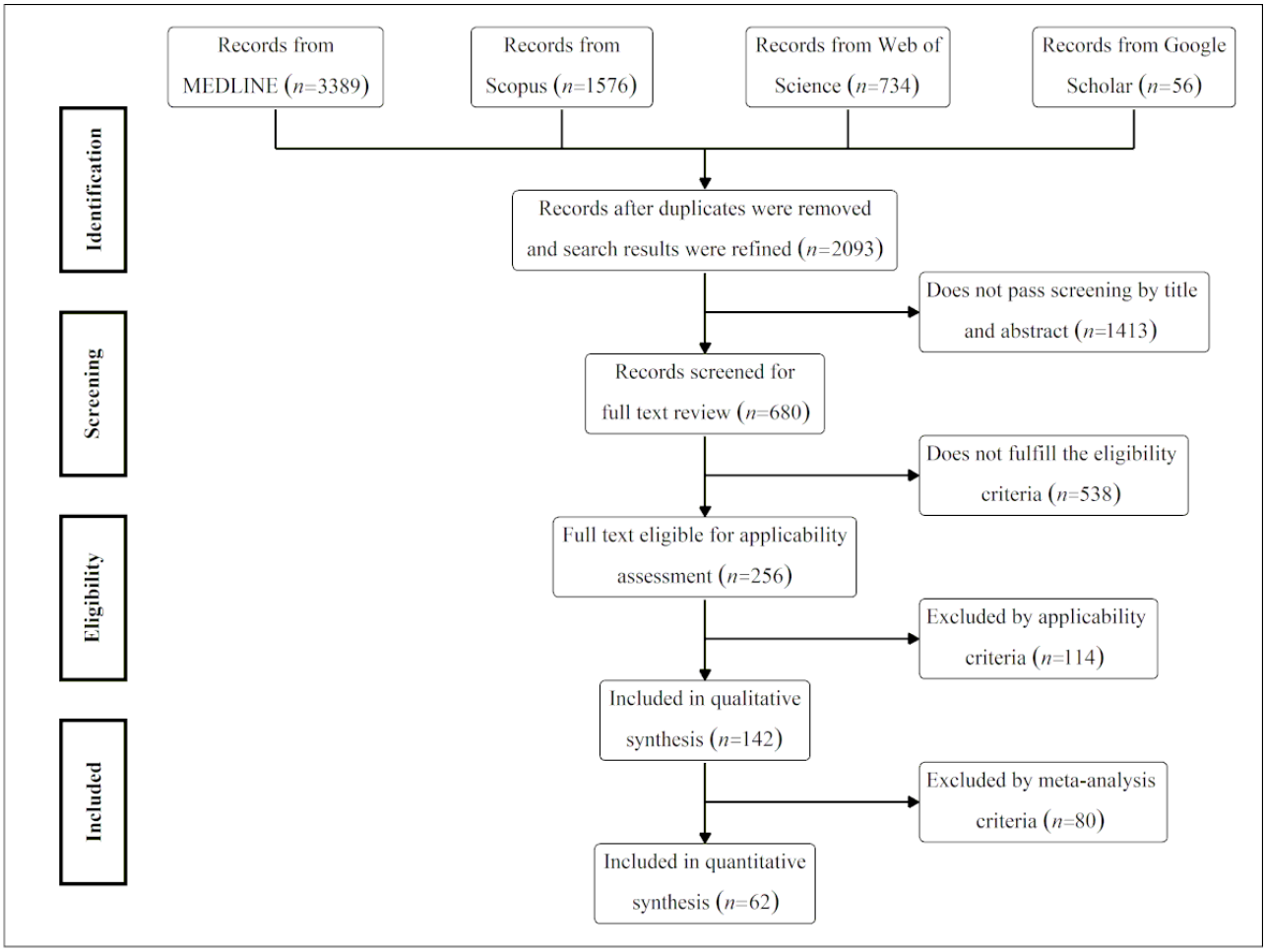
We described predictors in the prediction models from studies in the meta-analysis. For each outcome in the meta-analysis, we selected only random effects models in which an algorithm significantly outperformed the other. This was determined by the 95% CI of the difference in logit AUROCs between a non-LR and an LR model for an outcome. If any, we only selected those that included only non-LR low ROB studies. Only predictors in the final model were included. This was intended to elucidate predictor-outcome interactions that characterized an algorithm if it outperformed the others for a particular outcome.

Results

Study Selection

We found 2093 records from 4 literature databases (Figure 1). The search filters consisted of 144 combinations of keywords from 8 machine learning terms and 18 MeSH terms for pregnancy outcomes recursively derived from the keywords “machine learning AND pregnancy” (Multimedia Appendix 1). We refined the search results, identified research articles (not including conference abstracts or theses), and removed duplicates. After screening and eligibility assessment, we included 142 studies for the qualitative analysis, of which 62 were used for the quantitative analysis. A detailed description of the eligible criteria, process of study selection, and list of studies for the full-text review are given in Multimedia Appendix 1.

Figure 1. Study selection workflow.



Characteristics of the Studies

Briefly, we collected studies that either developed or validated a prediction model applying either LR (77/142, 54.2%) [29-105] or non-LR machine learning algorithms (50/142, 35.2%; Table 1) [106-155]. Overall, 15 studies applied both LR and non-LR algorithms (15/142, 10.6%) [156-170]. The cohort population of the studies in this review consisted of every type of population, study design, timing, and setting that we desired to discuss in this review. More studies discussed fetuses or newborns than pregnant women in non-LR prediction studies (26/50, 52% vs 11/50, 22%). Meanwhile, the opposite occurred in LR studies that focused more on pregnant women than fetuses or newborns (50/77, 65% vs 19/77, 25%). Most used data sets were from retrospective cohorts for LR (53/77, 69%) [29-36,38-42,47,49-54,56-61,64-66,68-71,76-80,83,87-90,92,94,95,97,100-105], non-LR prediction studies (27/50, 54%) [107,108,111-113,116,117,121,122,127,130-138,140,142,143,148,150,151,153,154], or both (9/15, 60%) [157-160,163-165,167-170]. A retrospective cohort is one of the recommended study designs for prognostic purposes instead of diagnostic prediction [21]. This corresponds to our review question that warrants prognostic predictions in pregnancy care intended to inform clinicians' decision making.

Only a few studies had prediction timing up to the puerperal or neonatal period for LR (2/77, 3%) [74,85], non-LR (3/50, 6%) [114,129,149], or both algorithms (2/15, 13%) [162,168]. This is because some predictors were assessed after delivery, whereas our review question demanded those be assessed up to delivery. We also considered studies using data sets from either primary care or hospital settings because the data are applicable for clinicians' decision making on a daily basis. As applicability was already included in the eligibility assessment before the qualitative analysis, eligible studies were not found to use data sets from either primary care or hospital settings, such as from a house-to-house survey or a screening program. Most used data sets were from hospital settings, whereas only a few of those were from primary care settings in the LR (6/77, 8%) [65,69,73,77,78,87], non-LR (6/50, 12%) [119,122,132,135,148,153], or both algorithms (1/15, 7%) [162]. A detailed description of this is also given in Multimedia Appendix 1.

Table 1. Characteristics of eligible studies.

Variable	Number of studies (percentage based on column total)			
	LR[a] (n=77), n (%)	Non-LR (n=50), n (%)	Both (n=15), n (%)	Total (n=142), n (%)
Population				
Pregnant women	50 (65)	11 (22)	6 (40)	67 (47.2)
Fetuses or newborns	19 (25)	26 (52)	7 (47)	52 (36.6)
Men or women in procreative management	8 (10)	13 (26)	2 (13)	23 (16.2)
Study design				
Retrospective	53 (69)	27 (54)	9 (60)	89 (62.7)
Nested case-control	4 (5)	14 (28)	2 (13)	20 (14.1)
Prospective	13 (17)	4 (8)	0 (0)	17 (12)
Cross-sectional	3 (4)	3 (6)	3 (20)	9 (6.3)
Case-control	4 (5)	2 (4)	1 (7)	7 (4.9)
Timing				
At delivery	28 (36)	26 (52)	7 (46.7)	61 (42.9)
At pregnancy	34 (44)	21 (42)	5 (33.3)	60 (42.3)
Mixed timing	13 (17)	0 (0)	1 (6.7)	14 (9.9)
Puerperal or neonatal period	2 (3)	3 (6)	2 (13.3)	7 (4.9)
Setting				
Hospital	61 (79)	43 (86)	9 (60)	113 (79.6)
Both	10 (13)	1 (2)	5 (33)	16 (11.3)
Primary care	6 (8)	6 (12)	1 (7)	13 (9.2)

[a]LR: logistic regression.

LR and Other Machine Learning Algorithms

Most studies applied an LR (92/142, 64.8%) to develop a prediction model (Table 2). Meanwhile, an artificial neural network was mostly applied by non-LR studies (20/142, 14.1%). Studies that applied LR and non-LR algorithms mostly compared LR with an artificial neural network (5/15, 33%) [161,163,165,166,170] and decision tree (5/15, 33%) [156,159,167-169], but decision trees tended to be paired with an LR compared with an artificial neural network (5/7, 71% vs 5/20, 25%).

The characteristics of study populations showed that pregnant women and fetuses or newborns were the populations of most studies developed using LR and non-LR models, respectively. Among pregnant women, the LR algorithm was mostly applied to develop predictions for outcome categories of obstetric labor (13/77, 17%) [36,46,47,54,57,62,64,70,83,86,91,97,103], pregnancy-induced hypertension (12/77, 16%) [30,31,43,48,55,65,66,68,76,81,93,105], and gestational diabetes (7/77, 9%) [33,45,49,84,94,100,104]. Among fetus or newborn populations, non-LR algorithms were mostly applied to develop predictions for outcome categories of premature birth (12/50, 24%) [111,112,115,116,118,119,121,122,125,130,141,143] and fetal distress (9/50, 18%) [113,124,128,137,138,145,146,152,155]. In addition, more non-LR algorithms (13/20, 65%) were applied for the outcome category of *in vitro* fertilization than for the LR algorithm.

Table 2. Machine learning algorithm and category of outcome.

Variable	Number of studies (percentage based on column total)			
	LR[a] (n=77), n (%)	Non-LR (n=50), n (%)	Both (n=15), n (%)	Total (n=142), n (%)
Machine learning algorithm				
Logistic regression	77 (100)	N/A[b]	15 (100)	92 (64.8)
Artificial neural network	N/A	15 (30)	5 (33)	20 (14.1)
Support vector machine	N/A	9 (18)	1 (7)	10 (7.0)
Deep neural network	N/A	8 (16)	1 (7)	9 (6.3)
Random forest	N/A	7 (14)	1 (7)	8 (5.6)
Decision tree	N/A	2 (4)	5 (33)	7 (4.9)
Gradient boosting	N/A	3 (6)	2 (13)	5 (3.5)
Naïve Bayes	N/A	4 (8)	0 (0)	4 (2.8)
Ensemble of algorithms	N/A	2 (4)	0 (0)	2 (1.4)
Category of outcome				
Premature birth	9 (12)	12 (24)	3 (20)	24 (16.9)
In vitro fertilization	7 (9)	13 (26)	2 (13)	22 (15.5)
Obstetric labor	13 (17)	1 (2)	2 (13)	16 (11.3)
Pregnancy-induced hypertension	12 (16)	4 (8)	0 (0)	16 (11.3)
Fetal distress	1 (1)	9 (18)	0 (0)	10 (7.0)
Gestational diabetes	7 (9)	2 (4)	1 (7)	10 (7.0)
Cesarean section	4 (5)	3 (6)	2 (13)	9 (6.3)
Fetal development	4 (5)	1 (2)	0 (0)	5 (3.5)
Small-for-gestational-age infant	3 (4)	1 (2)	1 (7)	5 (3.5)
Others	17 (22)	4 (8)	4 (27)	25 (17.6)

[a]LR: logistic regression.

[b]N/A: not applicable.

ROB Within and Across Studies

ROB is described for each eligible study in Multimedia Appendix 1 [29-170]. Among the 142 eligible studies, there were 24 (16.9%) low ROB studies [38,61-63,71,98,104,110,113,115,117-119,128,134,141,142,145,147,149,155,157,158,169], 117 (82.4%) high ROB studies [29-37,39-60,64-70,72-97,99-103,105-109,111,112,114,116,120-123,125-127,129-133,135-140,143,144,146,148,150-154,156,159-168,170], and 1 (0.7%) unclear ROB study (Table 3) [124]. Among the low ROB studies, the categories of outcomes were premature birth (7/24, 30%) [38,63,115,118,119,141,169], fetal distress (5/24, 21%) [71,113,128,145,155], *in vitro* fertilization (4/24, 17%) [61,110,134,158], gestational diabetes (2/24, 8%) [104,157], cesarean section (CS; 2/24, 8%) [117,142], obstetric labor (1/24, 4%) [62], pregnancy-induced hypertension (1/24, 4%) [147], central nervous system malformations (1/24, 4%) [149], and others (1/24, 4%) [98].

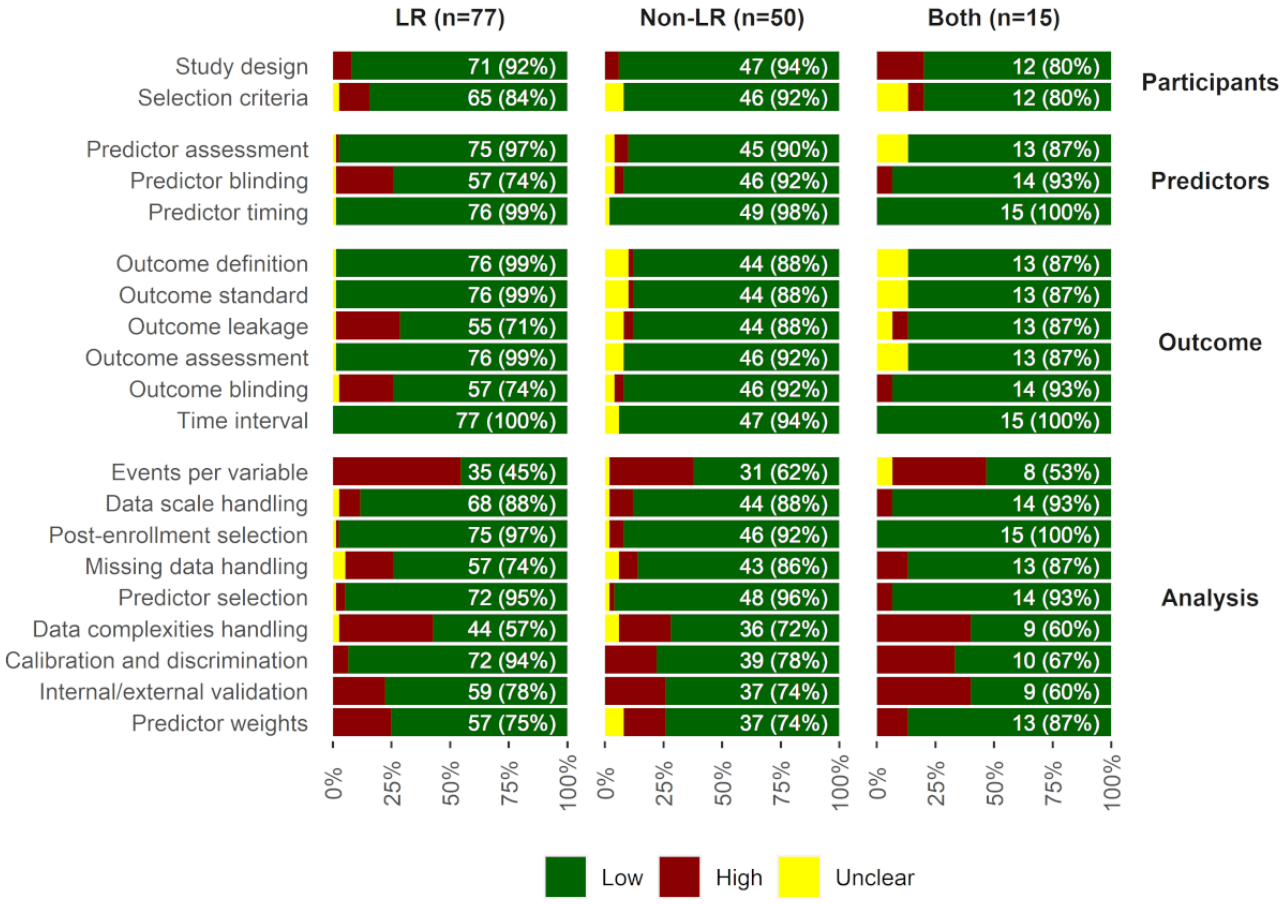
Table 3. Risk of bias within studies.

Assessment by domain	Studies by algorithm			
	LR[a] (n=77), n (%)	Non-LR (n=50), n (%)	Both (n=15), n (%)	Total (n=142), n (%)
Participants				
Low	60 (78)	44 (88)	11 (73)	115 (80.9)
High	15 (19)	3 (6)	4 (27)	22 (15.5)
Unclear	2 (3)	3 (6)	0 (0)	5 (3.5)
Predictors				
Low	54 (70)	43 (86)	12 (80)	109 (76.8)
High	20 (26)	5 (10)	1 (7)	26 (18.3)
Unclear	3 (4)	2 (4)	2 (13)	7 (4.9)
Outcome				
Low	51 (66)	40 (80)	11 (74)	102 (71.8)
High	24 (31)	4 (8)	2 (13)	30 (21.1)
Unclear	2 (3)	6 (12)	2 (13)	10 (7.1)
Analysis				
Low	8 (10)	15 (30)	3 (20)	26 (18.3)
High	69 (90)	35 (70)	12 (80)	116 (81.7)
Unclear	0 (0)	0 (0)	0 (0)	0 (0.0)
Overall				
Low	7 (9)	14 (28)	3 (20)	24 (16.9)
High	70 (91)	35 (70)	12 (80)	117 (82.4)
Unclear	0 (0)	1 (2)	0 (0)	1 (0.7)

[a]LR: logistic regression.

ROB is also described across the studies in Table 3 and Figure 2. The corresponding signaling questions for each term and the answers for each study are described in Multimedia Appendix 1. Low ROB studies were the fewest in the analysis domain (26/142, 18.3%), consisted of the LR (8/77, 10%) [38,61-64,71,96,98,104], non-LR (15/50, 30%) [63,110,113,115,117-119,124,128,134,141,142,145,147,155], and both algorithms (3/15, 20%) [157,158,169]. In the analysis domain, the fewest low ROB studies that achieved the minimum events per variable (EPV) consisted of LR (35/77, 45%) and non-LR (31/50, 62%) prediction studies. More calibration and discrimination tests were conducted using LR (72/77, 94%) than by non-LR (39/50, 78%) prediction studies. In contrast, more non-LR prediction studies appropriately handled missing data (43/50, 86%) compared with LR prediction studies (57/77, 74%).

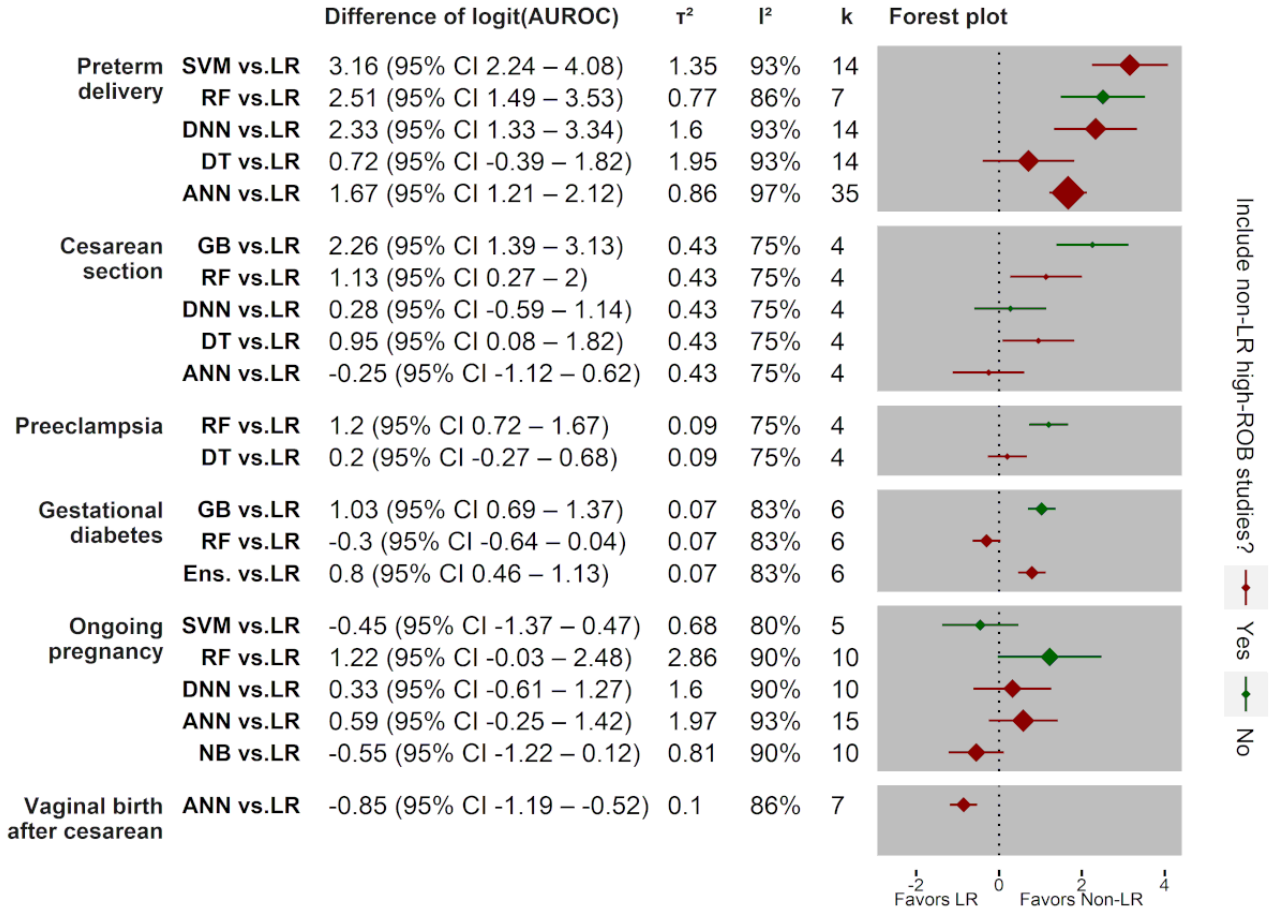
Figure 2. Signaling questions with respect to ROB domains across studies. Bars from low/high/unclear ROB are stacked to be 100%. Domains are described on the right-hand side. The number on the bar is the number of low ROB studies (total LR/non-LR/both at top) based on a single signaling question summarized as a term on the left-hand side. LR: logistic regression; ROB: risk of bias.



Comparison of the Predictive Performance

There were 62 studies in the meta-analysis that had outcomes that were predicted by at least one non-LR and 3 LR models (see *Summary Measures* section). Overall, 21 random effects models of the predictive performance by non-LR versus LR models are shown in a forest plot (Figure 3). Forest plots of logit AUROC differences for each random effects model are described (Multimedia Appendix 1). With respect to candidate studies (n) included in the final random effects models, we developed 5 random effects models for preterm delivery (20/62, 32%) [32,44,60,63,75,87,96,111,112,115,118,119,121,125,130,141,143,156,163,169], 5 for CS (7/62,11%) [79,90,106,117,142,166,167], 2 for pre-eclampsia (6/62, 10%) [31,48,65,76,123,147], 3 for gestational diabetes (9/62, 15%) [33,45,84,94,100,104,108,139,157], 5 for ongoing pregnancy (13/62, 21%) [73,78,99,110,132,134-136,148,150,153,158,170], and 1 for vaginal birth after CS (7/62, 11%) [36,47,57,64,83,97,165].

Figure 3. Forest plot of random effects models for differences in logit AUROCs from a non-LR with any LR prediction models. Plots were grouped by outcome. The lines indicate the 95% CI with diamonds whose sizes were determined by the number of pairwise comparisons (k). Absolute and relative values of between-study heterogeneities are denoted by τ2 and I2, respectively. Colors of the boxes and lines were determined based on the existence of high ROB studies among those using non-LR algorithms. ANN: artificial neural network; AUROC: area under the receiver operating characteristic curve; DNN: deep neural network; DT: decision tree; Ens: ensemble of multiple algorithms; GB: gradient boosting; LR: logistic regression; NB: naïve Bayes; RF: random forest; ROB: risk of bias; SVM: support vector machine.



To determine the final random effects model for each comparison, we identified studies that were responsible as the source of heterogeneity and removed those AUROCs from the random effects model. We excluded a non-LR [121] and an LR study [84] that developed a prediction model for preterm delivery and gestational diabetes, respectively. This is because their AUROCs were outliers compared with those for the same outcome and algorithm. We also excluded 3 LR studies [32,63,87]. In those studies, preterm delivery was defined as delivering within 1 to 2 weeks of preterm labor presentation. Meanwhile, the majority of studies for this outcome defined preterm delivery as that before 37 weeks of gestation.

The non-LR models significantly outperformed the LR models in preterm delivery (4/5 non-LR models), CS (3/5 non-LR models), pre-eclampsia (1/2 non-LR models), and gestational diabetes (2/3 non-LR models). From those that examined preterm delivery, a prediction model did not include a non-LR high ROB study [115] compared with those from 7 LR studies [32,44,60,63,75,87,96]. This model applied a random forest (differences in logit AUROC 2.51; 95% CI 1.49-3.53). The same algorithm was applied to a prediction model from a non-LR low ROB study in pre-eclampsia [147]. For random effects modeling, this model also significantly outperformed those from 4 LR studies (1.2, 95% CI 0.72-1.67) [31,48,65,76]. Meanwhile, prediction models from non-LR low ROB studies of Saleem et al [142] and Artzi et al [157] significantly outperformed those from the corresponding LR studies as an aggregate for CS (2.26, 95% CI 1.39-3.13) and gestational diabetes (1.03, 95% CI 0.69-1.37). Interestingly, the models were developed using a gradient boosting algorithm that used multiple decision trees similar to a random forest.

In contrast, a prediction model using a non-LR algorithm significantly underperformed compared with those using an LR in a random effects model (−0.85, 95% CI −1.19 to −0.52). This applied an artificial neural network to predict vaginal birth after a CS [165]. This model underperformed compared with those from 7 LR studies [36,47,57,64,83,97,165]. However, the non-LR study was a high ROB study.

A random effects model developed for comparison of artificial neural networks and LR to predict preterm delivery had the highest heterogeneity by I^2 (97%; k=35). This number means that 97% of the total variability among 35 data points of differences in logit AUROCs was caused by between-studies heterogeneity instead of sampling error within each study [171]. This is reasonable because a higher variance occurs with a larger number of comparisons within a random effects model. In contrast, a random effects model with the smallest number of

comparisons (k=4) also had the lowest heterogeneity by I^2 (75%). This random effects model was developed to analyze comparisons of non-LR and LR algorithms for either CS or pre-eclampsia. Nevertheless, a diverse target population and hyperparameter optimization conceivably caused the heterogeneity of the predictive performance, although the same outcome was predicted using the same data set and machine learning algorithm. The lowest I^2 in this meta-analysis remains classified as substantial heterogeneity instead of moderate or unimportant; thus, performing random effects instead of fixed effect modeling is recommended to address this issue [172].

However, I^2 only indicates that the difference in logit AUROCs substantively varies across studies but does not tell how much this metric varies [173]. To interpret the absolute heterogeneity for the difference in logit AUROCs, we needed to consider the observed AUROC of a non-LR model for each of the random effects models. The observed AUROCs were described for each of the original studies in this meta-analysis in Multimedia Appendix 1.

A random effects model developed for comparison of random forests and LR to predict ongoing pregnancy had the highest absolute value of heterogeneity (τ^2=2.86). In this random effects model, random forests were applied to develop predictions in 2 studies that reported AUROCs of 0.740 (95% CI 0.710-0.770) [158] and 0.9820 [134]. We simulated a sequence of logit AUROCs to identify equivalent differences in AUROCs to approximate a difference of the logit value in the random effects model (1.22, 95% CI −0.03 to 2.48). AUROC differences of 0.206 and 0.026 were equivalent to a difference in the logit AUROC of 0.91, compared with those aggregated from LR models for the random forest models of Blank et al [158] and Mirroshandel et al [134], respectively. Using τ^2, one can calculate the 95% prediction interval (PI) of the logit AUROC difference, as previously described [173]. This estimates the potential AUROC of the random forest to predict ongoing pregnancy with respect to an LR using different populations. For this random effects model, the 95% PI of the logit AUROC difference ranged from −4.75 to 7.19. This is equivalent to 0.257 lower and >0.73 higher than AUROCs of any LRs in the random effects model for the random forest model of Blank et al [158]. For the random forest model of Mirroshandel et al [134], the 95% PI was equivalent to 0.018 lower and 0.943 higher than the AUROCs of any LRs in the random effects model. This is a reasonably wide PI for the highest τ^2 in this meta-analysis, although the non-LR study had a low ROB. This is because ROB only reflects the risk of a predictive performance that differs from the true value of the training sample. However, the ROB does not reflect the difference if the predictive performance is compared with other samples across different populations.

For the random effects model with the lowest τ^2 and including a non-LR low ROB study, the random effects model had a logit AUROC difference of 1.03 (95% CI 0.69-1.37) for a prediction model of gestational diabetes using gradient boosting. The prediction study reported an AUROC of 0.875 (95% CI 0.868-0.885) [157]. The 95% PI of the logit AUROC difference estimated an equivalent AUROC that ranged from 0.0096 lower to 0.425 higher than the AUROCs of any LR in the random effects model. The gradient boosting model from this study is likely to outperform an LR to predict gestational diabetes.

In addition, we may need to know the τ^2 meaning for the random effects model with the highest I^2 and larger numbers of comparisons (k). This random effects model had an AUROC difference of 1.67 (95% CI 1.21-1.94; 95% PI −2.08 to 5.42; k=35) for a prediction model of preterm delivery using an artificial neural network. Overall, 5 non-LR studies were included in this random effects model. The remaining studies reported AUROCs of 0.88 [111], 0.94 [118], 0.945 [125], 0.9115 [163], and 0.911 (95% CI 0.862-0.960) [130]. Considering only the lowest (0.862) and highest (0.960) that covered all of the AUROCs, the artificial neural network model may have AUROCs of 0.119 lower and 0.864 higher than those of any LR. The AUROC interval was also as wide as that of the random effects model with the highest τ^2.

Descriptive Analysis of Predictors

A random effects model was selected for each outcome except for ongoing pregnancy, which fulfilled our criteria to describe the predictors. For each outcome in the meta-analysis, we selected random effects models in which either a non-LR algorithm significantly outperformed the LR or it was significantly underperformed by the LR. This was determined by the 95% CI of the difference in the logit AUROCs between the non-LR and LR models for an outcome. If any, we only selected those including only non-LR low ROB studies. The random effects models were random forest versus LR for preterm delivery, gradient boosting versus LR for CS, random forest versus LR for pre-eclampsia, gradient boosting versus LR for gestational diabetes, and artificial neural network versus LR for vaginal birth after a CS. As we only extracted the AUROC of either the best LR or non-LR model, only predictors and outcomes of that model were considered if there were multiple models for different subtypes of the outcome in a study.

For preterm delivery, Despotovic et al [115] developed a random forest model using a previously published standardized electrohysterogram (EHG) data set [174]. This data set was also used by other studies in this meta-analysis to predict the same outcome using different algorithms [118,125,130,141,143,169]. All predictors were features extracted from the multichannel EHG obtained at around 22 and 32 weeks of gestation to predict delivery after 39 and 34 to <37 weeks of gestation for term and preterm delivery, respectively. Compared with their counterparts, LR models used predictors consisting of maternal demographics or lifestyle [44,60,75,96,163], medical or obstetric histories [44,75,96,156,163], clinical predictors from obstetrical examinations [44,163], EHG [169], and biomarkers [75]. These were obtained before pregnancy [60,96,156,163], at 11 to 14 weeks of gestation [75], 18 to 34 weeks of gestation [44,163,169], or near events within 1 to 2 weeks [44]. The LR models were developed to predict preterm delivery at 20 to <37 weeks of gestation [44,75,96,163,169] and any delivery at <37 weeks of gestation (predictors could be taken before pregnancy) [60,156].

For CS, Saleem et al [142] developed a gradient boosting model using a previously published standardized cardiotocogram (CTG) data set [175]. This data set was also used by Fergus et al [117] in this meta-analysis to predict the same outcome using a deep neural network. All predictors were features extracted from the CTG data set obtained at first- and second-stage labor for a maximum of 90 min preceding delivery to predict a CS. Compared with their counterparts, LR models used predictors consisting of maternal characteristics [79,90,166], medical histories [167], obstetric histories [90,166,167], and clinical predictors from obstetric examinations [90,166,167], ultrasound measures [79], routine laboratory tests [90], and medications [90]. These were obtained before [90,166,167] and during pregnancy [79,90,166,167]. The LR models were developed to predict CS [166,167], emergency CS [79], and CS in pregnant women with gestational hypertension or mild pre-eclampsia at term [90].

For pre-eclampsia, Sufriyana et al [147] developed a random forest model that used a nationwide health insurance data set. The predictors consisted of maternal demographics and medical histories but excluded obstetric ones. These were obtained before and during pregnancy up to 2 days before the events (pre-eclampsia or eclampsia of any severity and timing). Meanwhile, the LR counterparts used maternal demographics or lifestyle [31,65,76], medical histories [31,65,76], obstetric histories [31,65,76], family histories [31,76], clinical or obstetric examinations [31,65], ultrasound measures [65], routine laboratory tests [76], and biomarkers [48,65]. These predictors were obtained before pregnancy [31], at 11 to 13 weeks of gestation [65], and at <20 weeks of gestation [48]. LR models were developed to predict pre-eclampsia of any severity and timing [31,48,65,76]. The predictors were taken before pregnancy, and this disorder occurs after 20 weeks of gestation by definition.

For gestational diabetes, Artzi et al [157] developed a gradient boosting model that used a nongovernmental, nationwide health care database. The predictors consisted of maternal demographics, medical histories, obstetric histories, clinical or obstetric examinations, routine laboratory tests, and medications. These predictors were obtained before pregnancy and up to 22 weeks of gestation to predict gestational diabetes diagnosed at 24 to 28 weeks of gestation. The LR counterparts used maternal demographics or lifestyle [33,100,104], medical histories [33], obstetric histories [104], family histories [33,45], clinical examinations [33], obstetric examinations [33], routine laboratory tests [33,45,94,100,104], medications, and biomarkers [33,45]. The predictor timing was 6 to 14 weeks of gestation [33,45,94,100,104] and >14 to 22 weeks of gestation [45,100,104]. Meanwhile, the outcome timing was 24 to 28 weeks of gestation [33,45,94,100,104].

For vaginal birth after a CS, Macones et al [165] developed an artificial neural network model that used a medical records database. The predictors used maternal characteristics, medical histories, obstetric histories, obstetric examinations, and labor procedures. These were obtained before pregnancy, during pregnancy, and at labor to predict successful vaginal birth after a CS. The LR counterparts used maternal characteristics [36,47,64,83,97], medical histories [57], obstetric histories [36,47,57,64,83,97], obstetric examinations [97], and labor procedures [97]. These were obtained before pregnancy [36,47,57,64,83,97], during pregnancy [97], and at labor [97]. The models predicted vaginal birth after a CS with the same definition as those of non-LR studies [36,47,57,64,83].

Discussion

Summary of Evidence

Of the 2093 records from 4 literature databases using 144 keywords, we found 142 eligible studies, among which 24 had a low ROB. These eligible studies developed prediction models for outcome categories of premature birth, *in vitro* fertilization, obstetric labor, pregnancy-induced hypertension, fetal distress, gestational diabetes, CS, fetal development, small-for-gestational-age infants, and others.

There were 4 models with non-LR algorithms from low ROB studies that had significantly higher differences in logit AUROCs than those with LR algorithms. The models used random forest algorithms to predict preterm delivery (2.51, 95% CI 1.49-3.53), gradient boosting algorithms to predict CS (2.26, 95% CI 1.39-3.13), random forest algorithms to predict pre-eclampsia (1.2, 95% CI 0.72-1.67), and gradient boosting algorithms to predict gestational diabetes (1.03, 95% CI 0.69-1.37). The first model that applied a random forest used only EHG records to predict preterm delivery. The second random forest model used only maternal demographics and medical histories but excluded obstetric ones for pre-eclampsia prediction. Meanwhile, the first model that applied a gradient boosting algorithm used only CTG records to predict CSs. The last model was developed by applying a gradient boosting algorithm for gestational diabetes. This model used maternal demographics, medical histories, obstetric histories, clinical or obstetric examinations, routine laboratory tests, and medications.

Comparisons With Prior Work

We compared our systematic review and meta-analysis with prior works related to either machine learning algorithms or pregnancy outcomes similar to those in our study. A recent paper described applications of artificial intelligence in obstetrics and gynecology [176]. That paper was a narrative instead of a scoping or systematic review. Our systematic review and meta-analysis covered all pregnancy outcomes in obstetrics, as described in that paper. These were described as fetal heart monitoring and pregnancy surveillance, gestational diabetes mellitus, preterm labor, parturition, and *in vitro* fertilization.

Nevertheless, the predicted outcomes by non-LR models in our review were still insufficient. Diseases that cause maternal deaths should receive higher priority than those causing neonatal deaths. The risks were higher for pregnant women with antepartum hemorrhage (incidence rate ratio [IRR]=3.5, 95% CI 2.0-6.1) or hypertension (IRR=1.5, 95% CI 1.1-2.2) compared with those without these diseases [177]. Maternal sepsis was also associated with fetal or neonatal deaths (odds ratio [OR] 5.78, 95% CI 2.89-11.21) [178]. Accordingly, the impact of the prediction models may be insufficient to reduce both maternal and neonatal deaths.

LR was found in our study to be the most often used algorithm to develop a prediction model in pregnancy care, including predicted outcomes that caused the most maternal deaths, followed by artificial (shallow) neural networks, support vector machines, and deep neural networks. These corresponded to a systematic review and meta-analysis [13] that showed a similar majority of machine learning algorithms in medicine, except that the study reported classification and regression trees to be the second most often used algorithms (30/71, 42%). All models within eligible studies in that review were included instead of only choosing the best one within each study. Using the same summary measures as we did, the aforementioned review demonstrated that non-LR models from low ROB studies did not outperform LR models. A decision tree showed a difference of logit AUROCs of −0.34 (95% CI −0.65 to −0.04; *k*=16) compared with an LR. The review selected 125 eligible studies of 927 candidates from one database. Between-study heterogeneity was not described in that review.

Similar to a previous study [13], a systematic review and meta-analysis did not consider LR as a machine learning algorithm and only compared the predictive performances of non-LR algorithms [179]. This study compared machine learning models to predict any outcomes using routinely collected intensive care unit data. Most of the algorithms were artificial neural networks (72/169, 42.6%), support vector machines (40/169, 23.7%), and decision trees (35/169, 20.7%). However, since 2015, most of the algorithms were support vector machines (37/125, 29.6%) and random forests (72/169, 42.6%). These corresponded to the majority of machine learning algorithms for pregnancy care in our systematic review.

We hold a particular assumption to determine whether interaction of predictors and outcome may be best predicted by a prediction algorithm. If the same predictors and outcomes were used by the best prediction algorithm applied in either non-LR or LR models but not used by the other outcomes in this meta-analysis, then the prediction algorithm may be the best for the pregnancy outcome using those predictors. To predict preterm delivery with predictors that included EHG in either non-LR or LR models [115,169], the random forest outperformed the LR algorithm. Similar to this model in terms of using biomedical signals, gradient boosting also outperformed LR using CTG [142], but none of the LR counterparts used the same predictor. Other predictors were used across outcomes and algorithms (LR or non-LR). These included maternal demographics, lifestyle, medical or obstetric histories, clinical examinations, ultrasound measures, routine laboratory tests, biomarkers, and medication or procedures. Family histories were used in the LR models to predict gestational diabetes in this meta-analysis but were not used by the gradient boosting model (the non-LR counterpart). Therefore, we could not find a convincing pattern of predictors with respect to the best algorithms for each of the other pregnancy outcomes beyond preterm delivery.

Interestingly, the random forest significantly outperformed the LR for almost all of the pregnancy outcomes included in the meta-analysis. Although the gradient boosting algorithm significantly outperformed the LR for CS and gestational diabetes instead of the random forest, gradient boosting also uses multiple decision trees as in the random forest. For ongoing pregnancy predictions in *in vitro* fertilization, a random forest model from low ROB studies also showed the largest difference in logit AUROCs outperforming LR (1.22, 95% CI −0.03 to 2.48) compared with other non-LR algorithms. For predicting vaginal delivery after a CS, a non-LR algorithm, particularly an artificial neural network in our meta-analysis, did not significantly outperform LR.

Comparing differences in AUROCs and focusing on multiple prediction algorithms, a study with individual participant data also compared LR and non-LR algorithms, particularly Poisson regression, random forest, gradient boosting, and an ensemble of a random forest with either LR or support vector machine [180]. Several models were developed to predict all-cause readmissions in patients with heart failure within 30 and 180 days. The random forest significantly outperformed the LR (0.601, 95% CI 0.594-0.607 vs 0.533, 95% CI 0.527-0.538) for 30-day readmissions. Similar to the random forest, the gradient boosting algorithm (0.613, 95% CI 0.607-0.618) also significantly outperformed the LR. The predictors consisted of medical histories and routine laboratory tests.

Massive evaluation of 179 algorithms from 17 machine learning families was conducted using 121 data sets [181]. The best results were achieved using random forests. In our review, there were 13 studies in which the best models applied either a random forest [106,108,115,134,144,147,155,158] or gradient boosting [127,140,142,157,160]. Random forests used multiple subsets of all samples and predictors randomly with replacement to grow multiple parallel decision trees [182]. Although gradient boosting also uses multiple decision trees, the advantages of random forest over gradient boosting are robust to noise and overfitting [183]. Meanwhile, gradient boosting randomly uses multiple subsets of all samples without replacement to sequentially construct additive regression models [184]. The advantages of gradient boosting over random forests are state-of-the-art predictive performance on tabular data and the customizability of loss of function [181,185]. Hence, several gradient boosting algorithms were developed, and some studies in our review applied these algorithms. To predict gestational diabetes, Artzi et al [157] applied LightGBM, a scalable gradient boosting machine. This algorithm was optimized to speed up the training process by up to 20-fold with the same accuracy [186]. Another gradient boosting system (ie, XGBoost) [187] was applied in a study by Qiu et al [140] to predict live births after *in vitro* fertilization. This study was not included in our meta-analysis because there was an insufficient number of LR [61,69] and gradient boosting [140] algorithms for predicting live births.

Of the pregnancy outcomes predicted by non-LR algorithms in this review, most outcomes were *in vitro* fertilization, premature birth, and fetal distress, possibly because of several reasons. Using keywords of “machine learning IVF” in MEDLINE, we found a review paper from 2011 call for a need for artificial intelligence in *in vitro* fertilization [188]. Only one machine learning study for *in vitro* fertilization was found before that study [189]. All machine learning studies for *in vitro* fertilization were published after the review paper, and most studies were identified within 2093 records in our review

[110,140,150,153,158,190-193]. As prediction for *in vitro* fertilization had already begun by 1989 [194], the machine learning prediction (non-LR) possibly arose because of the 2011 review. Meanwhile, for machine learning predictions of premature birth, fetal distress, and CS, many data sets (25/43, 58%) were secondary instead of primarily collected data. The secondary data sets consisted of predictors and outcomes of EHG and preterm delivery [174] (7/25, 28%), CTG, and acidotic blood pH of the umbilical artery [175] (4/25, 16%), CTG and CS [175] (2/25, 8%), CTG and acidotic blood pH of the umbilical artery [195] (3/25, 12%), EHG and preterm delivery [196] (2/25, 8%), and others (7/25, 28%). This implied that shared data sets drive more machine learning predictions compared with self-collected data sets. This indicates that the increase in publicly available data has driven progress in machine learning applications in health care [197].

For non-LR algorithms, the lack of shared data sets may have been the reason for few prediction studies for maternal outcomes compared with those for neonatal outcomes in this systematic review. Meanwhile, pregnancy-induced hypertension was found in pregnant women of newborns who were born prematurely [198]. Prematurity was also associated with maternal sepsis (OR 2.81, 95% CI 1.99-3.96), including antenatal cases [178]. Therefore, more shared data sets for maternal outcomes are needed. Future studies using machine learning algorithms should develop more prediction models for maternal outcomes in pregnancy care.

In addition, sample sizes of data sets for model development may contribute to bias in predictive performance. For example, in our meta-analysis, prediction models of ongoing pregnancy in *in vitro* fertilization had point estimates of AUROCs ranging from 0.575 to 0.982. These were developed using a support vector machine [110], artificial neural networks [132,136,170], random forests [134,158], deep neural networks [148,153], naïve Bayes algorithms [126,135,150], and LRs [73,78,99,158,170]. Compared with a recent systematic review focusing on prediction for *in vitro* fertilization [143,194], the range of AUROCs was wider than that of the previous review. The AUROCs ranged from 0.59 to 0.775 without non-LR machine learning predictions. A previous review also reported that the sample sizes ranged from 110 to 288,161 instances, whereas our review found that studies that applied non-LR algorithms alone or combined with LR had sample sizes ranging from only 46 [158] to 8836 [148] instances. Meanwhile, non-LR machine learning algorithms require larger sample sizes relative to the number of candidate predictors [199].

A meta-analysis of multivariable LR was also previously conducted for premature birth from 4 studies [200]. In a previous systematic review, the 2 highest AUROCs were 0.67 (95% CI 0.62-0.72; low ROB) and 0.64 (95% CI 0.60-0.68; high ROB). Non-LR models of premature birth in our systematic review showed AUROCs of 0.75 (95% CI 0.67-0.82) [121] and 0.911 (95% CI 0.862-0.96) [130], but these models were developed from high ROB studies. The other models only reported point estimates of the AUROC, which were a minimum of 0.6 by a decision tree [156] and a maximum of 0.991 by a support vector machine [143].

Minimizing the bias of model performance is the first thing to consider when developing a clinical prediction model. Several concerns need to be addressed when developing prognostic machine learning predictions of pregnancy care. In our review, most studies had problems of insufficient EPV (either LR and non-LR studies), single imputation (mostly LR studies), and no assessment of calibration (mostly non-LR studies). This may expose the studies to high ROBs [21]. The overestimation of the predictive performance is larger, with fewer participants with events relative to the number of predictor candidates, as described in the PROBAST guidelines. Most ROBs in our review were contributed by the domain of analysis, and answers to which the EPV signaling question mostly led studies to high ROB assessment results. Insufficient EPV mean that the study developed a model using a data set with a sample size that was less than the minimum requirement for events relative to the number of predictors. LR only requires 20 EPV, whereas non-LR algorithms require 50 to 200 EPV. Meanwhile, single imputation means that missing values are imputed by any random value, mean, median, mode, or one-time regression. Multiple imputations are more recommended than single imputations, in which the preferred method is multiple equations by chained equations. For the assessment of calibration, a study should show the incidence of events (true probability) for each subset of samples that belongs to the same range of predicted probability by the model. We recommend these based on PROBAST guidelines and other guidelines for machine learning prognostic predictions in pregnancy care [15,21].

Strengths and Limitations

Our systematic review and meta-analysis will allow investigators or clinicians in pregnancy care to consider whether trying multiple machine learning models provides benefit to their studies. If more prediction models are needed for the outcomes with more specific problems or subpopulations, then predictive modeling may consider comparisons of LR and non-LR algorithms for specific outcomes that were compared in our meta-analysis. We also reported heterogeneity measures to interpret the predictive performances of algorithms across studies.

However, the diverse populations and hyperparameters caused substantial heterogeneity of predictive performance in our meta-analysis. Future meta-analyses will be needed if more machine learning models are developed for the same outcome using the same algorithm. However, we tried to minimize the heterogeneity by excluding several studies to ensure more homogenous outcome definitions and normally distributed AUROCs. We also applied random effects modeling as recommended [172].

Conclusions

Prediction models using non-LR machine learning algorithms significantly outperformed those using LR for several pregnancy outcomes. These non-LR algorithms were random forests for predicting preterm delivery and pre-eclampsia and gradient boosting for predicting CS and gestational diabetes. In our review, studies that developed models using these algorithms had low ROBs. For predicting ongoing pregnancy in *in vitro* fertilization, non-LR algorithms did not significantly outperform

LR. Prediction models using non-LR algorithms for vaginal birth after a CS significantly underperformed LR, but the study with the non-LR algorithm had a high ROB.

On the basis of our meta-analysis, we recommend comparing multiple machine learning models, which include both LR and non-LR algorithms, to develop a prediction model. In our systematic review, we also found that many studies had high ROBs in the domain of analysis. In this domain, many studies lacked EPV to develop a prediction model. Hence, we also recommend the future development of a prediction model to pursue standard EPV and other standards based on guidelines to minimize ROBs.

Acknowledgments

This study was funded by the Ministry of Science and Technology of Taiwan under grant number MOST108-2221-E-038-018 and and MOST109-2221-E-038-018 to ES. The sponsor had no role in the research design or contents of the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details on forest plots, search filter, eligibility criteria, study selection, list of reviewed studies, risk of bias assessment, signaling questions and the answers, predictive performance and sample size, R code for meta-analysis, and records of studies.
[DOCX File , 2653 KB - medinform_v8i11e16503_app1.docx]

References

1. Domínguez-Almendros S, Benítez-Parejo N, Gonzalez-Ramirez A. Logistic regression models. Allergol Immunopathol (Madr) 2011;39(5):295-305. [doi: 10.1016/j.aller.2011.05.002] [Medline: 21820234]
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. Higgins JP. Nonlinear systems in medicine. Yale J Biol Med 2002;75(5-6):247-260 [FREE Full text] [Medline: 14580107]
4. The Millennium Development Goals Report. United Nations. 2015. URL: https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf [accessed 2019-07-25]
5. Say L, Chou D, Gemmill A, Tunçalp O, Moller A, Daniels J, et al. Global causes of maternal death: a WHO systematic analysis. Lancet Glob Health 2014 Jun;2(6):e323-e333 [FREE Full text] [doi: 10.1016/S2214-109X(14)70227-X] [Medline: 25103301]
6. Lehtonen L, Gimeno A, Parra-Llorca A, Vento M. Early neonatal death: a challenge worldwide. Semin Fetal Neonatal Med 2017 Jun;22(3):153-160. [doi: 10.1016/j.siny.2017.02.006] [Medline: 28238633]
7. Burlinson CE, Sirounis D, Walley KR, Chau A. Sepsis in pregnancy and the puerperium. Int J Obstet Anesth 2018 Nov;36:96-107. [doi: 10.1016/j.ijoa.2018.04.010] [Medline: 29921485]
8. Edwards HM. Aetiology and treatment of severe postpartum haemorrhage. Dan Med J 2018 Mar;65(3):- [FREE Full text] [Medline: 29510809]
9. Nair TM. Statistical and artificial neural network-based analysis to understand complexity and heterogeneity in preeclampsia. Comput Biol Chem 2018 Aug;75:222-230. [doi: 10.1016/j.compbiolchem.2018.05.011] [Medline: 29859381]
10. Romero R, Dey SK, Fisher SJ. Preterm labor: one syndrome, many causes. Science 2014 Aug 15;345(6198):760-765 [FREE Full text] [doi: 10.1126/science.1251816] [Medline: 25124429]
11. Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I. Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis. Asian Pac J Cancer Prev 2018 Jul 27;19(7):1747-1752 [FREE Full text] [doi: 10.22034/APJCP.2018.19.7.1747] [Medline: 30049182]
12. Lee Y, Ragguett R, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 2018 Dec 1;241:519-532. [doi: 10.1016/j.jad.2018.08.073] [Medline: 30153635]
13. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
14. Gregório T, Pipa S, Cavaleiro P, Atanásio G, Albuquerque I, Chaves PC, et al. Prognostic models for intracerebral hemorrhage: systematic review and meta-analysis. BMC Med Res Methodol 2018 Nov 20;18(1):145 [FREE Full text] [doi: 10.1186/s12874-018-0613-8] [Medline: 30458727]
15. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]

16. Nguyen AV, Blears EE, Ross E, Lall RR, Ortega-Barnett J. Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis. Neurosurg Focus 2018 Nov 1;45(5):E5. [doi: 10.3171/2018.8.FOCUS18325] [Medline: 30453459]
17. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
18. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
19. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744 [FREE Full text] [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
20. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015 Jan 6;162(1):55-63. [doi: 10.7326/M14-0697] [Medline: 25560714]
21. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]
22. Mitchell TM. Machine Learning. New York, NY: McGraw-Hill Inc; 1997.
23. James SB, Bardenet R, Bengio Y, Balázs K. Algorithms for Hyper-Parameter Optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011 Presented at: NIPS'11; December 21-24, 2011; Granada, Spain URL: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf [doi: 10.5555/2986459.2986743]
24. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. J Clin Epidemiol 2016 Jan;69:40-50 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.05.009] [Medline: 26142114]
25. Cheung M, Ho R, Lim Y, Mak A. Conducting a meta-analysis: basics and good practices. Int J Rheum Dis 2012 Apr;15(2):129-135. [doi: 10.1111/j.1756-185X.2012.01712.x] [Medline: 22462415]
26. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. Br Med J 2011 Feb 10;342:d549. [doi: 10.1136/bmj.d549] [Medline: 21310794]
27. Viechtbauer W. Conducting meta-analyses in with the package. J Stat Soft 2010;36(3):-. [doi: 10.18637/jss.v036.i03]
28. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods 2016 Mar;7(1):55-79 [FREE Full text] [doi: 10.1002/jrsm.1164] [Medline: 26332144]
29. Allouche M, Huissoud C, Guyard-Boileau B, Rouzier R, Parant O. Development and validation of nomograms for predicting preterm delivery. Am J Obstet Gynecol 2011 Mar;204(3):242.e1-242.e8. [doi: 10.1016/j.ajog.2010.09.030] [Medline: 21093847]
30. Almeida S, Katz L, Coutinho I, Amorim M. Validation of fullPIERS model for prediction of adverse outcomes among women with severe pre-eclampsia. Int J Gynaecol Obstet 2017 Aug;138(2):142-147. [doi: 10.1002/ijgo.12197] [Medline: 28475234]
31. Al-Rubaie ZT, Hudson HM, Jenkins G, Mahmoud I, Ray JG, Askie LM, et al. Prediction of pre-eclampsia in nulliparous women using routinely collected maternal characteristics: a model development and validation study. BMC Pregnancy Childbirth 2020 Jan 6;20(1):23 [FREE Full text] [doi: 10.1186/s12884-019-2712-x] [Medline: 31906891]
32. Bastek JA, Sammel MD, Srinivas SK, McShea MA, Foreman MN, Elovitz MA, et al. Clinical prediction rules for preterm birth in patients presenting with preterm labor. Obstet Gynecol 2012 Jun;119(6):1119-1128. [doi: 10.1097/AOG.0b013e31825503e5] [Medline: 22617575]
33. Benhalima K, van Crombrugge P, Moyson C, Verhaeghe J, Vandeginste S, Verlaenen H, et al. Estimating the risk of gestational diabetes mellitus based on the 2013 WHO criteria: a prediction model based on clinical and biochemical variables in early pregnancy. Acta Diabetol 2020 Jun;57(6):661-671. [doi: 10.1007/s00592-019-01469-5] [Medline: 31915927]
34. Berntorp K, Anderberg E, Claesson R, Ignell C, Källén K. The relative importance of maternal body mass index and glucose levels for prediction of large-for-gestational-age births. BMC Pregnancy Childbirth 2015 Oct 29;15:280 [FREE Full text] [doi: 10.1186/s12884-015-0722-x] [Medline: 26514116]
35. Broekmans FJ, Verweij PJ, Eijkemans MJ, Mannaerts BM, Witjes H. Prognostic models for high and low ovarian responses in controlled ovarian stimulation using a GnRH antagonist protocol. Hum Reprod 2014 Aug;29(8):1688-1697 [FREE Full text] [doi: 10.1093/humrep/deu090] [Medline: 24903202]
36. Fagerberg MC, Källén K. Third-trimester prediction of successful vaginal birth after one cesarean delivery-a Swedish model. Acta Obstet Gynecol Scand 2020 May;99(5):660-668. [doi: 10.1111/aogs.13783] [Medline: 31788783]

37. Casikar I, Lu C, Reid S, Condous G. Prediction of successful expectant management of first trimester miscarriage: development and validation of a new mathematical model. Aust N Z J Obstet Gynaecol 2013 Feb;53(1):58-63. [doi: 10.1111/ajo.12053] [Medline: 23405997]
38. Cerqueira FR, Ferreira TG, de Paiva Oliveira A, Augusto DA, Krempser E, Corrêa Barbosa HJ, et al. NICeSim: an open-source simulator based on machine learning techniques to support medical research on prenatal and perinatal care decision making. Artif Intell Med 2014 Nov;62(3):193-201. [doi: 10.1016/j.artmed.2014.10.001] [Medline: 25457563]
39. Chandrasekaran S, Bastek JA, Turitz AL, Durnwald CP. A prediction score to assess the risk of delivering a large for gestational age infant among obese women. J Matern Fetal Neonatal Med 2016;29(1):22-26. [doi: 10.3109/14767058.2014.991709] [Medline: 25428834]
40. Chen L, Luo D, Yu X, Jin M, Cai W. Predicting stress urinary incontinence during pregnancy: combination of pelvic floor ultrasound parameters and clinical factors. Acta Obstet Gynecol Scand 2018 Aug;97(8):966-975 [FREE Full text] [doi: 10.1111/aogs.13368] [Medline: 29754393]
41. Ciobanu A, Rouvali A, Syngelaki A, Akolekar R, Nicolaides KH. Prediction of small for gestational age neonates: screening by maternal factors, fetal biometry, and biomarkers at 35-37 weeks' gestation. Am J Obstet Gynecol 2019 May;220(5):486.e1-486.11. [doi: 10.1016/j.ajog.2019.01.227] [Medline: 30707967]
42. Cortet M, Maucort-Boulch D, Deneux-Tharaux C, Dupont C, Rudigoz R, Roy P, et al. Severity of post-partum hemorrhage after vaginal delivery is not predictable from clinical variables available at the time post-partum hemorrhage is diagnosed. J Obstet Gynaecol Res 2015 Feb;41(2):199-206. [doi: 10.1111/jog.12528] [Medline: 25303234]
43. Crovetto F, Figueras F, Triunfo S, Crispi F, Rodriguez-Sureda V, Dominguez C, et al. First trimester screening for early and late preeclampsia based on maternal characteristics, biophysical parameters, and angiogenic factors. Prenat Diagn 2015 Feb;35(2):183-191. [doi: 10.1002/pd.4519] [Medline: 25346181]
44. de Oliveira RV, Martins MD, Rios LT, Araujo Júnior E, Simões VM, Nardozza LM, et al. Predictive model for spontaneous preterm labor among pregnant women with contractions and intact amniotic membranes. Arch Gynecol Obstet 2012 Oct;286(4):893-900. [doi: 10.1007/s00404-012-2397-0] [Medline: 22674420]
45. de Wilde MA, Veltman-Verhulst SM, Goverde AJ, Lambalk CB, Laven JS, Franx A, et al. Preconception predictors of gestational diabetes: a multicentre prospective cohort study on the predominant complication of pregnancy in polycystic ovary syndrome. Hum Reprod 2014 Jun;29(6):1327-1336. [doi: 10.1093/humrep/deu077] [Medline: 24777850]
46. Eggebø TM, Wilhelm-Benartzi C, Hassan WA, Usman S, Salvesen KA, Lees CC. A model to predict vaginal delivery in nulliparous women based on maternal characteristics and intrapartum ultrasound. Am J Obstet Gynecol 2015 Sep;213(3):362.e1-362.e6. [doi: 10.1016/j.ajog.2015.05.044] [Medline: 26008180]
47. Fagerberg MC, Maršál K, Källén K. Predicting the chance of vaginal delivery after one cesarean section: validation and elaboration of a published prediction model. Eur J Obstet Gynecol Reprod Biol 2015 May;188:88-94. [doi: 10.1016/j.ejogrb.2015.02.031] [Medline: 25801723]
48. Guo Z, Yang F, Zhang J, Zhang Z, Li K, Tian Q, et al. Whole-genome promoter profiling of plasma DNA exhibits diagnostic value for placenta-origin pregnancy complications. Adv Sci (Weinh) 2020 Apr;7(7):1901819 [FREE Full text] [doi: 10.1002/advs.201901819] [Medline: 32274292]
49. Harper LM, Glover AV, Biggio JR, Tita A. Predicting failure of glyburide therapy in gestational diabetes. J Perinatol 2016 May;36(5):347-351 [FREE Full text] [doi: 10.1038/jp.2015.216] [Medline: 26796130]
50. Isono W, Nagamatsu T, Uemura Y, Fujii T, Hyodo H, Yamashita T, et al. Prediction model for the incidence of emergent cesarean section during induction of labor specialized in nulliparous low-risk women. J Obstet Gynaecol Res 2011 Dec;37(12):1784-1791. [doi: 10.1111/j.1447-0756.2011.01607.x] [Medline: 21793999]
51. Kang J, Kim HS, Lee EB, Uh Y, Han K, Park EY, et al. Prediction model for massive transfusion in placenta previa during cesarean section. Yonsei Med J 2020 Feb;61(2):154-160 [FREE Full text] [doi: 10.3349/ymj.2020.61.2.154] [Medline: 31997624]
52. Kawakita T, Mokhtari N, Huang JC, Landy HJ. Evaluation of risk-assessment tools for severe postpartum hemorrhage in women undergoing cesarean delivery. Obstet Gynecol 2019 Dec;134(6):1308-1316. [doi: 10.1097/AOG.0000000000003574] [Medline: 31764744]
53. Khan N, Ciobanu A, Karampitsakos T, Akolekar R, Nicolaides KH. Prediction of large-for-gestational-age neonate by routine third-trimester ultrasound. Ultrasound Obstet Gynecol 2019 Sep;54(3):326-333. [doi: 10.1002/uog.20377] [Medline: 31236963]
54. Kok M, van der Steeg J, van der Post J, Mol B. Prediction of success of external cephalic version after 36 weeks. Am J Perinatol 2011 Feb;28(2):103-110. [doi: 10.1055/s-0030-1262909] [Medline: 20661845]
55. Lafalla O, Esteban LM, Lou AC, Cornudella R, Domínguez M, Sanz G, et al. Clinical utility of thrombophilia, anticoagulant treatment, and maternal variables as predictors of placenta-mediated pregnancy complications: an extensive analysis. J Matern Fetal Neonatal Med 2019 May 9:1-11. [doi: 10.1080/14767058.2019.1611764] [Medline: 31018724]
56. Lee JS, Sultana R, Han NL, Sia AT, Sng BL. Development and validation of a predictive risk factor model for epidural re-siting in women undergoing labour epidural analgesia: a retrospective cohort study. BMC Anesthesiol 2018 Nov 29;18(1):176 [FREE Full text] [doi: 10.1186/s12871-018-0638-x] [Medline: 30497401]

57. Mardy AH, Ananth CV, Grobman WA, Gyamfi-Bannerman C. A prediction model of vaginal birth after cesarean in the preterm period. Am J Obstet Gynecol 2016 Oct;215(4):513.e1-513.e7. [doi: 10.1016/j.ajog.2016.05.039] [Medline: 27262971]
58. McCowan LM, Thompson JM, Taylor RS, Baker PN, North RA, Poston L, SCOPE consortium. Prediction of small for gestational age infants in healthy nulliparous women using clinical and ultrasound risk factors combined with early pregnancy biomarkers. PLoS One 2017;12(1):e0169311 [FREE Full text] [doi: 10.1371/journal.pone.0169311] [Medline: 28068394]
59. McCowan LM, Thompson JM, Taylor RS, North RA, Poston L, Baker PN, SCOPE Consortium. Clinical prediction in early pregnancy of infants small for gestational age by customised birthweight centiles: findings from a healthy nulliparous cohort. PLoS One 2013;8(8):e70917 [FREE Full text] [doi: 10.1371/journal.pone.0070917] [Medline: 23940665]
60. Mehta-Lee SS, Palma A, Bernstein PS, Lounsbury D, Schlecht NF. A preconception nomogram to predict preterm delivery. Matern Child Health J 2017 Jan;21(1):118-127. [doi: 10.1007/s10995-016-2100-3] [Medline: 27461021]
61. Meijerink A, Cissen M, Mochtar M, Fleischer K, Thoonen I, de Melker A, et al. Prediction model for live birth in ICSI using testicular extracted sperm. Hum Reprod 2016 Sep;31(9):1942-1951. [doi: 10.1093/humrep/dew146] [Medline: 27406949]
62. Meister M, Cahill A, Conner S, Woolfolk C, Lowder J. Predicting obstetric anal sphincter injuries in a modern obstetric population. Am J Obstet Gynecol 2016 Sep;215(3):310.e1-310.e7. [doi: 10.1016/j.ajog.2016.02.041] [Medline: 26902989]
63. Menon R, Bhat G, Saade GR, Spratt H. Multivariate adaptive regression splines analysis to predict biomarkers of spontaneous preterm birth. Acta Obstet Gynecol Scand 2014 Apr;93(4):382-391. [doi: 10.1111/aogs.12344] [Medline: 24461165]
64. Metz TD, Stoddard GJ, Henry E, Jackson M, Holmgren C, Esplin S. Simple, validated vaginal birth after cesarean delivery prediction model for use at the time of admission. Obstet Gynecol 2013 Sep;122(3):571-578 [FREE Full text] [doi: 10.1097/AOG.0b013e31829f8ced] [Medline: 23921867]
65. Murtoniemi K, Villa PM, Matomäki J, Keikkala E, Vuorela P, Hämäläinen E, et al. Prediction of pre-eclampsia and its subtypes in high-risk cohort: hyperglycosylated human chorionic gonadotropin in multivariate models. BMC Pregnancy Childbirth 2018 Jul 3;18(1):279 [FREE Full text] [doi: 10.1186/s12884-018-1908-9] [Medline: 29970026]
66. Myers J, Kenny L, McCowan L, Chan E, Dekker G, Poston L, SCOPE consortium. Angiogenic factors combined with clinical risk factors to predict preterm pre-eclampsia in nulliparous women: a predictive test accuracy study. BJOG 2013 Sep;120(10):1215-1223. [doi: 10.1111/1471-0528.12195] [Medline: 23906160]
67. Oates J, Casikar I, Campain A, Müller S, Yang J, Reid S, et al. A prediction model for viability at the end of the first trimester after a single early pregnancy evaluation. Aust N Z J Obstet Gynaecol 2013 Feb;53(1):51-57. [doi: 10.1111/ajo.12046] [Medline: 23405996]
68. Payne BA, Groen H, Ukah UV, Ansermino JM, Bhutta Z, Grobman W, miniPIERS working group. Development and internal validation of a multivariable model to predict perinatal death in pregnancy hypertension. Pregnancy Hypertens 2015 Oct;5(4):315-321 [FREE Full text] [doi: 10.1016/j.preghy.2015.08.006] [Medline: 26597747]
69. Pettersson G, Andersen AN, Broberg P, Arce J. Pre-stimulation parameters predicting live birth after IVF in the long GnRH agonist protocol. Reprod Biomed Online 2010 May;20(5):572-581. [doi: 10.1016/j.rbmo.2010.02.014] [Medline: 20236862]
70. Pettersson K, Yousaf K, Ranstam J, Westgren M, Ajne G. Predictive value of traction force measurement in vacuum extraction: development of a multivariate prognostic model. PLoS One 2017;12(3):e0171938 [FREE Full text] [doi: 10.1371/journal.pone.0171938] [Medline: 28257459]
71. Ramanah R, Omar S, Guillien A, Pugin A, Martin A, Riethmuller D, et al. Predicting umbilical artery pH during labour: development and validation of a nomogram using fetal heart rate patterns. Eur J Obstet Gynecol Reprod Biol 2018 Jun;225:166-171. [doi: 10.1016/j.ejogrb.2018.04.008] [Medline: 29727787]
72. Reid S, Lu C, Condous G. Can we improve the prediction of pouch of Douglas obliteration in women with suspected endometriosis using ultrasound-based models? A multicenter prospective observational study. Acta Obstet Gynecol Scand 2015 Dec;94(12):1297-1306. [doi: 10.1111/aogs.12779] [Medline: 26399692]
73. Rinaudo P, Shen S, Hua J, Qian S, Prabhu U, Garcia E, et al. (1)H NMR based profiling of spent culture media cannot predict success of implantation for day 3 human embryos. J Assist Reprod Genet 2012 Dec;29(12):1435-1442 [FREE Full text] [doi: 10.1007/s10815-012-9877-9] [Medline: 23090745]
74. Ryu A, Cho NJ, Kim YS, Lee EY. Predictive value of serum uric acid levels for adverse perinatal outcomes in preeclampsia. Medicine (Baltimore) 2019 May;98(18):e15462. [doi: 10.1097/MD.0000000000015462] [Medline: 31045822]
75. Sananes N, Meyer N, Gaudineau A, Aissi G, Boudier E, Fritz G, et al. Prediction of spontaneous preterm delivery in the first trimester of pregnancy. Eur J Obstet Gynecol Reprod Biol 2013 Nov;171(1):18-22. [doi: 10.1016/j.ejogrb.2013.07.042] [Medline: 24012451]
76. Sandström A, Snowden JM, Höijer J, Bottai M, Wikström AK. Clinical risk assessment in early pregnancy for preeclampsia in nulliparous women: a population based cohort study. PLoS One 2019;14(11):e0225716 [FREE Full text] [doi: 10.1371/journal.pone.0225716] [Medline: 31774875]
77. Scheinhardt M, Lerman T, König IR, Griesinger G. Performance of prognostic modelling of high and low ovarian response to ovarian stimulation for IVF. Hum Reprod 2018 Aug 01;33(8):1499-1505. [doi: 10.1093/humrep/dey236] [Medline: 30007353]

78. Shi W, Zhang S, Zhao W, Xia X, Wang M, Wang H, et al. Factors related to clinical pregnancy after vitrified-warmed embryo transfer: a retrospective and multivariate logistic regression analysis of 2313 transfer cycles. Hum Reprod 2013 Jul;28(7):1768-1775. [doi: 10.1093/humrep/det094] [Medline: 23599130]
79. Sovio U, Smith GC. Blinded ultrasound fetal biometry at 36 weeks and risk of emergency Cesarean delivery in a prospective cohort study of low-risk nulliparous women. Ultrasound Obstet Gynecol 2018 Jul;52(1):78-86. [doi: 10.1002/uog.17513] [Medline: 28452133]
80. Stamatopoulos N, Lu C, Casikar I, Reid S, Mongelli M, Hardy N, et al. Prediction of subsequent miscarriage risk in women who present with a viable pregnancy at the first early pregnancy scan. Aust N Z J Obstet Gynaecol 2015 Oct;55(5):464-472. [doi: 10.1111/ajo.12395] [Medline: 26294017]
81. Stott D, Bolten M, Salman M, Paraschiv D, Douiri A, Kametas NA. A prediction model for the response to oral labetalol for the treatment of antenatal hypertension. J Hum Hypertens 2017 Feb;31(2):126-131. [doi: 10.1038/jhh.2016.50] [Medline: 27465979]
82. Stroux L, Redman CW, Georgieva A, Payne SJ, Clifford GD. Doppler-based fetal heart rate analysis markers for the detection of early intrauterine growth restriction. Acta Obstet Gynecol Scand 2017 Nov;96(11):1322-1329. [doi: 10.1111/aogs.13228] [Medline: 28862738]
83. Tessmer-Tuck JA, El-Nashar SA, Racek AR, Lohse CM, Famuyide AO, Wick MJ. Predicting vaginal birth after cesarean section: a cohort study. Gynecol Obstet Invest 2014;77(2):121-126 [FREE Full text] [doi: 10.1159/000357757] [Medline: 24525697]
84. Thériault S, Giguère Y, Massé J, Girouard J, Forest J. Early prediction of gestational diabetes: a practical model combining clinical and biochemical markers. Clin Chem Lab Med 2016 Mar;54(3):509-518. [doi: 10.1515/cclm-2015-0537] [Medline: 26351946]
85. Timmerman E, Oude Rengerink K, Pajkrt E, Opmeer BC, van der Post JA, Bilardo CM. Ductus venosus pulsatility index measurement reduces the false-positive rate in first-trimester screening. Ultrasound Obstet Gynecol 2010 Dec;36(6):661-667. [doi: 10.1002/uog.7706] [Medline: 20521242]
86. Tsur A, Batsry L, Toussia-Cohen S, Rosenstein MG, Barak O, Brezinov Y, et al. Development and validation of a machine-learning model for prediction of shoulder dystocia. Ultrasound Obstet Gynecol 2020 Oct;56(4):588-596. [doi: 10.1002/uog.21878] [Medline: 31587401]
87. van Baaren GJ, Bruijn MM, Vis JY, Wilms FF, Oudijk MA, Kwee A, et al. Risk factors for preterm delivery: do they add to fetal fibronectin testing and cervical length measurement in the prediction of preterm delivery in symptomatic women? Eur J Obstet Gynecol Reprod Biol 2015 Sep;192:79-85. [doi: 10.1016/j.ejogrb.2015.05.004] [Medline: 26182836]
88. Van Calster B, Condous G, Kirk E, Bourne T, Timmerman D, Van Huffel S. An application of methods for the probabilistic three-class classification of pregnancies of unknown location. Artif Intell Med 2009 Jun;46(2):139-154. [doi: 10.1016/j.artmed.2008.12.003] [Medline: 19157812]
89. van der Ham DP, van Kuijk S, Opmeer BC, Willekes C, van Beek JJ, Mulder AL, PPROMEXIL trial group. Can neonatal sepsis be predicted in late preterm premature rupture of membranes? Development of a prediction model. Eur J Obstet Gynecol Reprod Biol 2014 May;176:90-95. [doi: 10.1016/j.ejogrb.2014.02.003] [Medline: 24630296]
90. van der Tuuk K, van Pampus MG, Koopmans C, Aarnoudse J, van den Berg PP, van Beek JJ, HYPITAT study group. Prediction of cesarean section risk in women with gestational hypertension or mild preeclampsia at term. Eur J Obstet Gynecol Reprod Biol 2015 Aug;191:23-27. [doi: 10.1016/j.ejogrb.2015.05.009] [Medline: 26070123]
91. Verhoeven CJ, Nuij C, Janssen-Rolf CR, Schuit E, Bais JM, Oei SG, et al. Predictors for failure of vacuum-assisted vaginal delivery: a case-control study. Eur J Obstet Gynecol Reprod Biol 2016 May;200:29-34. [doi: 10.1016/j.ejogrb.2016.02.008] [Medline: 26967343]
92. Vieira MC, White SL, Patel N, Seed PT, Briley AL, Sandall J, UPBEAT Consortium. Prediction of uncomplicated pregnancies in obese women: a prospective multicentre study. BMC Med 2017 Nov 3;15(1):194 [FREE Full text] [doi: 10.1186/s12916-017-0956-8] [Medline: 29096631]
93. Visentin S, Londero AP, Camerin M, Grisan E, Cosmi E. A possible new approach in the prediction of late gestational hypertension: the role of the fetal aortic intima-media thickness. Medicine (Baltimore) 2017 Jan;96(2):e5515. [doi: 10.1097/MD.0000000000005515] [Medline: 28079791]
94. Wang C, Zhu W, Wei Y, Su R, Feng H, Lin L, et al. The predictive effects of early pregnancy lipid profiles and fasting glucose on the risk of gestational diabetes mellitus stratified by body mass index. J Diabetes Res 2016;2016:3013567. [doi: 10.1155/2016/3013567] [Medline: 26981541]
95. Wang L, Matsunaga S, Mikami Y, Takai Y, Terui K, Seki H. Pre-delivery fibrinogen predicts adverse maternal or neonatal outcomes in patients with placental abruption. J Obstet Gynaecol Res 2016 Jul;42(7):796-802. [doi: 10.1111/jog.12988] [Medline: 27075198]
96. Weber A, Darmstadt GL, Gruber S, Foeller ME, Carmichael SL, Stevenson DK, et al. Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women. Ann Epidemiol 2018 Nov;28(11):783-9.e1. [doi: 10.1016/j.annepidem.2018.08.008] [Medline: 30236415]
97. Xing Y, Qi X, Wang X, Yang F. Development of a modified score system as prediction model for successful vaginal birth after cesarean delivery. Clin Transl Sci 2019 Jan;12(1):53-57 [FREE Full text] [doi: 10.1111/cts.12603] [Medline: 30548202]

98. Xu H, Feng G, Wei Y, Feng Y, Yang R, Wang L, et al. Predicting ectopic pregnancy using human chorionic gonadotropin (HCG) levels and main cause of infertility in women undergoing assisted reproductive treatment: retrospective observational cohort study. JMIR Med Inform 2020 Apr 16;8(4):e17366 [FREE Full text] [doi: 10.2196/17366] [Medline: 32297865]
99. Xu H, Wei Y, Yang R, Feng G, Tang W, Zhang H, et al. Prospective observational cohort study: computational models for early prediction of ongoing pregnancy in fresh IVF/ICSI-ET protocols. Life Sci 2019 Apr 1;222:221-227. [doi: 10.1016/j.lfs.2019.03.012] [Medline: 30858125]
100. Yang H, Zhu C, Ma Q, Long Y, Cheng Z. Variations of blood cells in prediction of gestational diabetes mellitus. J Perinat Med 2015 Jan;43(1):89-93. [doi: 10.1515/jpm-2014-0007] [Medline: 24897392]
101. Yang T, Li N, Qiao C, Liu C. Development of a novel nomogram for predicting placenta accreta in patients with scarred uterus: a retrospective cohort study. Front Med (Lausanne) 2019;6:289. [doi: 10.3389/fmed.2019.00289] [Medline: 31921868]
102. Yu C, Zhang R, Li J. A predictive model for high-quality blastocyst based on blastomere number, fragmentation, and symmetry. J Assist Reprod Genet 2018 May;35(5):809-816 [FREE Full text] [doi: 10.1007/s10815-018-1132-6] [Medline: 29502189]
103. Zhao R, Zhang W, Zhou L, Chen Y. Building a predictive model for successful vaginal delivery in nulliparas with term cephalic singleton pregnancies using decision tree analysis. J Obstet Gynaecol Res 2019 Aug;45(8):1536-1544. [doi: 10.1111/jog.14011] [Medline: 31161703]
104. Zheng T, Ye W, Wang X, Li X, Zhang J, Little J, et al. A simple model to predict risk of gestational diabetes mellitus from 8 to 20 weeks of gestation in Chinese women. BMC Pregnancy Childbirth 2019 Jul 19;19(1):252 [FREE Full text] [doi: 10.1186/s12884-019-2374-8] [Medline: 31324151]
105. Zwertbroek E, Broekhuijsen K, Langenveld J, van Baaren G, van den Berg P, Bremer H, HYPITAT-II Study Group. Prediction of progression to severe disease in women with late preterm hypertensive disorders of pregnancy. Acta Obstet Gynecol Scand 2017 Jan;96(1):96-105. [doi: 10.1111/aogs.13051] [Medline: 27792243]
106. Abbas SA, Riaz R, Kazmi SZ, Rizvi SS, Kwon SJ. Cause analysis of caesarian sections and application of machine learning methods for classification of birth data. IEEE Access 2018;6(5):67555-67561. [doi: 10.1109/ACCESS.2018.2879115]
107. Alberola-Rubio J, Garcia-Casado J, Prats-Boluda G, Ye-Lin Y, Desantes D, Valero J, et al. Prediction of labor onset type: spontaneous vs induced; role of electrohysterography? Comput Methods Programs Biomed 2017 Jun;144:127-133. [doi: 10.1016/j.cmpb.2017.03.018] [Medline: 28494996]
108. Balani J, Hyer S, Shehata H, Mohareb F. Visceral fat mass as a novel risk factor for predicting gestational diabetes in obese pregnant women. Obstet Med 2018 Sep;11(3):121-125 [FREE Full text] [doi: 10.1177/1753495X17754149] [Medline: 30214477]
109. Benalcazar-Parra C, Ye-Lin Y, Garcia-Casado J, Monfort-Ortiz R, Alberola-Rubio J, Perales A, et al. Prediction of labor induction success from the uterine electrohysterogram. J Sensors 2019 Nov 15;2019:1-12. [doi: 10.1155/2019/6916251]
110. Borup R, Thuesen L, Andersen C, Nyboe-Andersen A, Ziebe S, Winther O, et al. Competence classification of cumulus and granulosa cell transcriptome in embryos matched by morphology and female age. PLoS One 2016;11(4):e0153562 [FREE Full text] [doi: 10.1371/journal.pone.0153562] [Medline: 27128483]
111. Chen L, Hao Y. Feature extraction and classification of EHG between pregnancy and labour group using Hilbert-Huang transform and extreme learning machine. Comput Math Methods Med 2017;2017:7949507. [doi: 10.1155/2017/7949507] [Medline: 28316639]
112. Chen L, Hao Y, Hu X. Detection of preterm birth in electrohysterogram signals based on wavelet transform and stacked sparse autoencoder. PLoS One 2019;14(4):e0214712 [FREE Full text] [doi: 10.1371/journal.pone.0214712] [Medline: 30990810]
113. Cömert Z, Kocamaz AF, Subha V. Prognostic model based on image-based time-frequency features and genetic algorithm for fetal hypoxia assessment. Comput Biol Med 2018 Aug 1;99:85-97. [doi: 10.1016/j.compbiomed.2018.06.003] [Medline: 29894897]
114. Coppedè F, Grossi E, Migheli F, Migliore L. Polymorphisms in folate-metabolizing genes, chromosome damage, and risk of Down syndrome in Italian women: identification of key factors using artificial neural networks. BMC Med Genomics 2010 Sep 24;3:42 [FREE Full text] [doi: 10.1186/1755-8794-3-42] [Medline: 20868477]
115. Despotovic D, Zec A, Mladenovic K, Radin N, Turukalo T. A machine learning approach for an early prediction of preterm delivery. In: 16th International Symposium on Intelligent Systems and Informatics. 2018 Presented at: SISY'18; September 13-15, 2018; Subotica, Serbia. [doi: 10.1109/SISY.2018.8524818]
116. Elaveyini U, Devi SP, Rao KS. Neural networks prediction of preterm delivery with first trimester bleeding. Arch Gynecol Obstet 2011 May;283(5):971-979. [doi: 10.1007/s00404-010-1469-2] [Medline: 20449599]
117. Fergus P, Hussain A, Al-Jumeily D, Huang D, Bouguila N. Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms. Biomed Eng Online 2017 Jul 6;16(1):89 [FREE Full text] [doi: 10.1186/s12938-017-0378-z] [Medline: 28679415]
118. Fergus P, Idowu I, Hussain A, Dobbins C. Advanced artificial neural network classification for detecting preterm births using EHG records. Neurocomputing 2016 May;188:42-49. [doi: 10.1016/j.neucom.2015.01.107]

119. Fergus P, Montanez A, Abdulaimma B, Lisboa P, Chalmers C, Pineles B. Utilizing deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American women. IEEE/ACM Trans Comput Biol Bioinform 2020;17(2):668-678. [doi: 10.1109/TCBB.2018.2868667] [Medline: 30183645]
120. Figueras F, Savchev S, Triunfo S, Crovetto F, Gratacos E. An integrated model with classification criteria to predict small-for-gestational-age fetuses at risk of adverse perinatal outcome. Ultrasound Obstet Gynecol 2015 Mar;45(3):279-285. [doi: 10.1002/uog.14714] [Medline: 25358519]
121. Fiset S, Martel A, Glanc P, Barrett J, Melamed N. Prediction of spontaneous preterm birth among twin gestations using machine learning and texture analysis of cervical ultrasound images. Univ Tor Med J 2019;96(1):6-9 [FREE Full text]
122. Gao C, Osmundson S, Velez Edwards DR, Jackson G, Malin B, Chen Y. Deep learning predicts extreme preterm birth from electronic health records. J Biomed Inform 2019 Dec;100:103334. [doi: 10.1016/j.jbi.2019.103334] [Medline: 31678588]
123. Garcés MF, Sanchez E, Cardona LF, Simanca EL, González I, Leal LG, et al. Maternal serum meteorin levels and the risk of preeclampsia. PLoS One 2015;10(6):e0131013 [FREE Full text] [doi: 10.1371/journal.pone.0131013] [Medline: 26121675]
124. Georgoulas G, Karvelis P, Spilka J, Chudáček V, Stylios CD, Lhotská L. Investigating pH based evaluation of fetal heart rate (FHR) recordings. Health Technol (Berl) 2017;7(2):241-254 [FREE Full text] [doi: 10.1007/s12553-017-0201-7] [Medline: 29201590]
125. Hamdi M, Limem M, Maaref M. Detection and classification of nonstationary signals: application to uterine EMG for prognostication of premature delivery. Neurophysiology 2019 Dec 4;51(4):272-280. [doi: 10.1007/s11062-019-09821-9]
126. Hernández-González J, Inza I, Crisol-Ortíz L, Guembe M, Iñarra MJ, Lozano J. Fitting the data from embryo implantation prediction: learning from label proportions. Stat Methods Med Res 2018 Apr;27(4):1056-1066. [doi: 10.1177/0962280216651098] [Medline: 27242336]
127. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang S, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One 2019;14(8):e0221202 [FREE Full text] [doi: 10.1371/journal.pone.0221202] [Medline: 31442238]
128. Leonarduzzi R, Spilka J, Frecon J, Wendt H, Pustelnik N, Jaffard S, et al. P-leader multifractal analysis and sparse SVM for intrapartum fetal acidosis detection. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:1971-1974. [doi: 10.1109/EMBC.2015.7318771] [Medline: 26736671]
129. Li H, Luo M, Zheng J, Luo J, Zeng R, Feng N, et al. An artificial neural network prediction model of congenital heart disease based on risk factors: a hospital-based case-control study. Medicine (Baltimore) 2017 Feb;96(6):e6090. [doi: 10.1097/MD.0000000000006090] [Medline: 28178169]
130. Mas-Cabo J, Prats-Boluda G, Garcia-Casado J, Alberola-Rubio J, Perales A, Ye-Lin Y. Design and assessment of a robust and generalizable ANN-based classifier for the prediction of premature birth by means of multichannel electrohysterographic records. J Sensors 2019 Nov 25;2019:1-13. [doi: 10.1155/2019/5373810]
131. Mello G, Parretti E, Ognibene A, Mecacci F, Cioni R, Scarselli G, et al. Prediction of the development of pregnancy-induced hypertensive disorders in high-risk pregnant women by artificial neural networks. Clin Chem Lab Med 2001 Sep;39(9):801-805. [doi: 10.1515/CCLM.2001.132] [Medline: 11601676]
132. Milewski R, Kuczyńska A, Stankiewicz B, Kuczyński W. How much information about embryo implantation potential is included in morphokinetic data? A prediction model based on artificial neural networks and principal component analysis. Adv Med Sci 2017 Mar;62(1):202-206. [doi: 10.1016/j.advms.2017.02.001] [Medline: 28384614]
133. Milewski R, Milewska AJ, Więsak T, Morgan A. Comparison of artificial neural networks and logistic regression analysis in pregnancy prediction using the in vitro fertilization treatment. Stud Log Gramm Rhetor 2013;35((1)):39-48. [doi: 10.2478/slgr-2013-0033]
134. Mirroshandel S, Ghasemian F, Monji-Azad S. Applying data mining techniques for increasing implantation rate by selecting best sperms for intra-cytoplasmic sperm injection treatment. Comput Methods Programs Biomed 2016 Dec;137:215-229. [doi: 10.1016/j.cmpb.2016.09.013] [Medline: 28110726]
135. Morales DA, Bengoetxea E, Larrañaga P, García M, Franco Y, Fresnada M, et al. Bayesian classification for the selection of in vitro human embryos using morphological and clinical data. Comput Methods Programs Biomed 2008 May;90(2):104-116. [doi: 10.1016/j.cmpb.2007.11.018] [Medline: 18190996]
136. Paydar K, Niakan Kalhori SR, Akbarian M, Sheikhtaheri A. A clinical decision support system for prediction of pregnancy outcome in pregnant women with systemic lupus erythematosus. Int J Med Inform 2017 Jan;97:239-246. [doi: 10.1016/j.ijmedinf.2016.10.018] [Medline: 27919382]
137. Petrozziello A, Jordanov I, Aris Papageorghiou T, Christopher Redman WG, Georgieva A. Deep learning for continuous electronic fetal monitoring in labor. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:5866-5869. [doi: 10.1109/EMBC.2018.8513625] [Medline: 30441670]
138. Petrozziello A, Redman C, Papageorghiou A, Jordanov I, Georgieva A. Multimodal convolutional neural networks to detect fetal compromise during labor and delivery. IEEE Access 2019;7:112026-112036. [doi: 10.1109/access.2019.2933368]
139. Qiu H, Yu H, Wang L, Yao Q, Wu S, Yin C, et al. Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy. Sci Rep 2017 Nov 27;7(1):16417. [doi: 10.1038/s41598-017-16665-y] [Medline: 29180800]

140. Qiu J, Li P, Dong M, Xin X, Tan J. Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method. J Transl Med 2019 Sep 23;17(1):317 [FREE Full text] [doi: 10.1186/s12967-019-2062-5] [Medline: 31547822]
141. Sadi-Ahmed N, Kacha B, Taleb H, Kedir-Talha M. Relevant features selection for automatic prediction of preterm deliveries from pregnancy electrohysterograhic (EHG) records. J Med Syst 2017 Nov 11;41(12):204. [doi: 10.1007/s10916-017-0847-8] [Medline: 29128973]
142. Saleem S, Naqvi S, Manzoor T, Saeed A, Ur Rehman N, Mirza J. A strategy for classification of 'vaginal vs cesarean section' delivery: bivariate empirical mode decomposition of cardiotocographic recordings. Front Physiol 2019;10:246. [doi: 10.3389/fphys.2019.00246] [Medline: 30941054]
143. Shahbakhti M, Beiramvand M, Bavi M, Mohammadi Far S. A new efficient algorithm for prediction of preterm labor. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:4669-4672. [doi: 10.1109/EMBC.2019.8857837] [Medline: 31946904]
144. Signorini MG, Pini N, Malovini A, Bellazzi R, Magenes G. Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. Comput Methods Programs Biomed 2020 Mar;185:105015. [doi: 10.1016/j.cmpb.2019.105015] [Medline: 31678794]
145. Spilka J, Frecon J, Leonarduzzi R, Pustelnik N, Abry P, Doret M. Intrapartum fetal heart rate classification from trajectory in Sparse SVM feature space. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:2335-2338. [doi: 10.1109/EMBC.2015.7318861] [Medline: 26736761]
146. Spilka J, Frecon J, Leonarduzzi R, Pustelnik N, Abry P, Doret M. Sparse support vector machine for intrapartum fetal heart rate classification. IEEE J Biomed Health Inform 2017 May;21(3):664-671. [doi: 10.1109/JBHI.2016.2546312] [Medline: 27046884]
147. Sufriyana H, Wu Y, Su EC. Artificial intelligence-assisted prediction of preeclampsia: development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. EBioMedicine 2020 Apr;54:102710 [FREE Full text] [doi: 10.1016/j.ebiom.2020.102710] [Medline: 32283530]
148. Tran D, Cooke S, Illingworth P, Gardner D. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Hum Reprod 2019 Jun 4;34(6):1011-1018 [FREE Full text] [doi: 10.1093/humrep/dez064] [Medline: 31111884]
149. Troisi J, Landolfi A, Sarno L, Richards S, Symes S, Adair D, et al. A metabolomics-based approach for non-invasive screening of fetal central nervous system anomalies. Metabolomics 2018 May 25;14(6):77. [doi: 10.1007/s11306-018-1370-8] [Medline: 30830338]
150. Uyar A, Bener A, Ciray HN. Predictive modeling of implantation outcome in an in vitro fertilization setting: an application of machine learning methods. Med Decis Making 2015 Aug;35(6):714-725. [doi: 10.1177/0272989X14535984] [Medline: 24842951]
151. Uyar A, Bener A, Ciray H. ROC Based Evaluation and Comparison of Classifiers for IVF Implantation Prediction. In: International Conference on Electronic Healthcare. 2009 Presented at: eHealth'09; September 23-25, 2009; Istanbul, Turkey. [doi: 10.1007/978-3-642-11745-9_17]
152. Valensise H, Facchinetti F, Vasapollo B, Giannini F, Monte ID, Arduini D. The computerized fetal heart rate analysis in post-term pregnancy identifies patients at risk for fetal distress in labour. Eur J Obstet Gynecol Reprod Biol 2006 Apr 1;125(2):185-192. [doi: 10.1016/j.ejogrb.2005.06.034] [Medline: 16459010]
153. VerMilyea M, Hall J, Diakiw S, Johnston A, Nguyen T, Perugini D, et al. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Hum Reprod 2020 Apr 28;35(4):770-784 [FREE Full text] [doi: 10.1093/humrep/deaa013] [Medline: 32240301]
154. Vogiatzi P, Pouliakis A, Siristatidis C. An artificial neural network for the prediction of assisted reproduction outcome. J Assist Reprod Genet 2019 Jul;36(7):1441-1448 [FREE Full text] [doi: 10.1007/s10815-019-01498-7] [Medline: 31218565]
155. Xu L, Georgieva A, Redman C, Payne S. Feature selection for computerized fetal heart rate analysis using genetic algorithms. Annu Int Conf IEEE Eng Med Biol Soc 2013;2013:445-448. [doi: 10.1109/EMBC.2013.6609532] [Medline: 24109719]
156. Amini P, Maroufizadeh S, Samani RO, Hamidi O, Sepidarkish M. Prevalence and determinants of preterm birth in Tehran, Iran: a comparison between logistic regression and decision tree methods. Osong Public Health Res Perspect 2017 Jun;8(3):195-200 [FREE Full text] [doi: 10.24171/j.phrp.2017.8.3.06] [Medline: 28781942]
157. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben-Haroush A, et al. Prediction of gestational diabetes based on nationwide electronic health records. Nat Med 2020 Jan;26(1):71-76. [doi: 10.1038/s41591-019-0724-8] [Medline: 31932807]
158. Blank C, Wildeboer R, DeCroo I, Tilleman K, Weyers B, de Sutter P, et al. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. Fertil Steril 2019 Feb;111(2):318-326. [doi: 10.1016/j.fertnstert.2018.10.030] [Medline: 30611557]
159. Isakov O, Reicher L, Lavie A, Yogev Y, Maslovitz S. Prediction of success in external cephalic version for breech presentation at term. Obstet Gynecol 2019 May;133(5):857-866. [doi: 10.1097/AOG.0000000000003196] [Medline: 30969207]
160. Koivu A, Sairanen M. Predicting risk of stillbirth and preterm pregnancies with machine learning. Health Inf Sci Syst 2020 Dec;8(1):14 [FREE Full text] [doi: 10.1007/s13755-020-00105-9] [Medline: 32226625]

161. Kuhle S, Maguire B, Zhang H, Hamilton D, Allen AC, Joseph KS, et al. Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. BMC Pregnancy Childbirth 2018 Aug 15;18(1):333 [FREE Full text] [doi: 10.1186/s12884-018-1971-2] [Medline: 30111303]
162. Kumar SN, Saxena P, Patel R, Sharma A, Pradhan D, Singh H, et al. Predicting risk of low birth weight offspring from maternal features and blood polycyclic aromatic hydrocarbon concentration. Reprod Toxicol 2020 Jun;94:92-100. [doi: 10.1016/j.reprotox.2020.03.009] [Medline: 32283251]
163. Lee K, Ahn KH. Artificial neural network analysis of spontaneous preterm labor and birth and its major determinants. J Korean Med Sci 2019 Apr 29;34(16):e128 [FREE Full text] [doi: 10.3346/jkms.2019.34.e128] [Medline: 31020816]
164. Liu B, Shi S, Wu Y, Thomas D, Symul L, Pierson E, et al. Predicting pregnancy using large-scale data from a women's health tracking mobile application. Proc Int World Wide Web Conf 2019 May;2019:2999-3005 [FREE Full text] [doi: 10.1145/3308558.3313512] [Medline: 31538145]
165. Macones GA, Hausman N, Edelstein R, Stamilio DM, Marder SJ. Predicting outcomes of trials of labor in women attempting vaginal birth after cesarean delivery: a comparison of multivariate methods with neural networks. Am J Obstet Gynecol 2001 Feb;184(3):409-413. [doi: 10.1067/mob.2001.109386] [Medline: 11228495]
166. Maroufizadeh S, Amini P, Hosseini M, Almasi-Hashiani A, Mohammadi M, Navid B, et al. Determinants of cesarean section among primiparas: a comparison of classification methods. Iran J Public Health 2018 Dec;47(12):1913-1922 [FREE Full text] [Medline: 30788307]
167. Sims CJ, Meyn L, Caruana R, Rao R, Mitchell T, Krohn M. Predicting cesarean delivery with decision tree models. Am J Obstet Gynecol 2000 Nov;183(5):1198-1206. [doi: 10.1067/mob.2000.108891] [Medline: 11084566]
168. Agopian A, Lupo P, Tinker S, Canfield M, Mitchell L, National Birth Defects Prevention Study. Working towards a risk prediction model for neural tube defects. Birth Defects Res A Clin Mol Teratol 2012 Mar;94(3):141-146 [FREE Full text] [doi: 10.1002/bdra.22883] [Medline: 22253139]
169. Fergus P, Cheung P, Hussain A, Al-Jumeily D, Dobbins C, Iram S. Prediction of preterm deliveries from EHG signals using machine learning. PLoS One 2013;8(10):e77154 [FREE Full text] [doi: 10.1371/journal.pone.0077154] [Medline: 24204760]
170. Wald M, Sparks A, Sandlow J, Van-Voorhis B, Syrop C, Niederberger C. Computational models for prediction of IVF/ICSI outcomes with surgically retrieved spermatozoa. Reprod Biomed Online 2005 Sep;11(3):325-331. [doi: 10.1016/s1472-6483(10)60840-1] [Medline: 16176672]
171. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods 2006 Jun;11(2):193-206. [doi: 10.1037/1082-989X.11.2.193] [Medline: 16784338]
172. Deeks J, Higgins J, Altman D. Analysing data undertaking meta-analyses. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. New Jersey, USA: Wiley-Blackwell; 2008:243-296.
173. Borenstein M, Higgins JP, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods 2017 Mar;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
174. Fele-Zorz G, Kavsek G, Novak-Antolic Z, Jager F. A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups. Med Biol Eng Comput 2008 Sep;46(9):911-922. [doi: 10.1007/s11517-008-0350-y] [Medline: 18437439]
175. Chudáček V, Spilka J, Burša M, Janků P, Hruban L, Huptych M, et al. Open access intrapartum CTG database. BMC Pregnancy Childbirth 2014 Jan 13;14:16 [FREE Full text] [doi: 10.1186/1471-2393-14-16] [Medline: 24418387]
176. Iftikhar P, Kuijpers M, Khayyat A, Iftikhar A, DeGouvia de Sa M. A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups. Cureus 2020 Feb 28;12(2):e7124 [FREE Full text] [doi: 10.7759/cureus.7124] [Medline: 32257670]
177. Khanam R, Ahmed S, Creanga AA, Begum N, Koffi AK, Mahmud A, Projahnmo Study Group in Bangladesh. Antepartum complications and perinatal mortality in rural Bangladesh. BMC Pregnancy Childbirth 2017 Mar 7;17(1):81 [FREE Full text] [doi: 10.1186/s12884-017-1264-1] [Medline: 28270117]
178. Knowles SJ, O'Sullivan NP, Meenan AM, Hanniffy R, Robson M. Maternal sepsis incidence, aetiology and outcome for mother and fetus: a prospective study. BJOG 2015 Apr;122(5):663-671. [doi: 10.1111/1471-0528.12892] [Medline: 24862293]
179. Shillan D, Sterne JA, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care 2019 Aug 22;23(1):284 [FREE Full text] [doi: 10.1186/s13054-019-2564-9] [Medline: 31439010]
180. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li S, et al. Analysis of machine learning techniques for heart failure readmissions. Circ Cardiovasc Qual Outcomes 2016 Nov;9(6):629-640 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.116.003039] [Medline: 28263938]
181. Fernandez-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 2014;15(90):3133-3181 [FREE Full text]
182. Breiman L. Random Forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
183. Fawagreh K, Gaber M, Elyan E. Random forests: from early developments to recent advancements. Syst Sci Control Engin 2014 Oct 06;2(1):602-609. [doi: 10.1080/21642583.2014.956265]

184. Friedman J. Stochastic gradient boosting. Comput Stat Data Analy 2002 Feb;38(4):367-378. [doi: 10.1016/s0167-9473(01)00065-2]
185. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21. [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
186. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 1-7, 2017; Long Beach URL: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf
187. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: ACM'16; August 11-15, 2016; San Francisco, California, USA. [doi: 10.1145/2939672.2939785]
188. Siristatidis C, Pouliakis A, Chrelias C, Kassanos D. Artificial intelligence in IVF: a need. Syst Biol Reprod Med 2011 Aug;57(4):179-185. [doi: 10.3109/19396368.2011.558607] [Medline: 21375363]
189. Haake KW, List P, Baier D, Zimmermann G, Pretzsch G, Alexander H. [Risk assessment in ovarian hyperstimulation syndrome (OHS) using the machine learning system (Decision Master) in 155 in-vitro fertilisations and embryo-transfer (IVF/ET) cycles with a long stimulation protocol]. Zentralbl Gynakol 1997;119(Suppl 1):23-27. [Medline: 9245120]
190. Manna C, Nanni L, Lumini A, Pappalardo S. Artificial intelligence techniques for embryo and oocyte classification. Reprod Biomed Online 2013 Jan;26(1):42-49. [doi: 10.1016/j.rbmo.2012.09.015] [Medline: 23177416]
191. Santos Filho E, Noble J, Poli M, Griffiths T, Emerson G, Wells D. A method for semi-automatic grading of human blastocyst microscope images. Hum Reprod 2012 Sep;27(9):2641-2648. [doi: 10.1093/humrep/des219] [Medline: 22736327]
192. Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ Digit Med 2019;2:21. [doi: 10.1038/s41746-019-0096-y] [Medline: 31304368]
193. Liang B, Gao Y, Xu J, Song Y, Xuan L, Shi T, et al. Raman profiling of embryo culture medium to identify aneuploid and euploid embryos. Fertil Steril 2019 Apr;111(4):753-62.e1. [doi: 10.1016/j.fertnstert.2018.11.036] [Medline: 30683589]
194. Ratna M, Bhattacharya S, Abdulrahim B, McLernon D. A systematic review of the quality of clinical prediction models in in vitro fertilisation. Hum Reprod 2020 Jan 1;35(1):100-116. [doi: 10.1093/humrep/dez258] [Medline: 31960915]
195. Doret M, Massoud M, Constans A, Gaucherand P. Use of peripartum ST analysis of fetal electrocardiogram without blood sampling: a large prospective cohort study. Eur J Obstet Gynecol Reprod Biol 2011 May;156(1):35-40. [doi: 10.1016/j.ejogrb.2010.12.042] [Medline: 21257256]
196. Alexandersson A, Steingrimsdottir T, Terrien J, Marque C, Karlsson B. The Icelandic 16-electrode electrohysterogram database. Sci Data 2015;2:150017. [doi: 10.1038/sdata.2015.17] [Medline: 25984349]
197. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
198. Kennady G, Kottarathara MJ, Kottarathara AJ, Ajith R, Anandakesavan TM, Ambujam K. Maternal and neonatal outcomes in pregnancy induced hypertension: an observational study. Clin Exp Obstet Gynecol 2017;44(1):110-112. [Medline: 29714877]
199. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014 Dec 22;14:137 [FREE Full text] [doi: 10.1186/1471-2288-14-137] [Medline: 25532820]
200. Meertens LJ, van Montfort P, Scheepers HC, van Kuijk SM, Aardenburg R, Langenveld J, et al. Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation. Acta Obstet Gynecol Scand 2018 Aug;97(8):907-920. [doi: 10.1111/aogs.13358] [Medline: 29663314]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CHARMS: checklist for critical appraisal and data extraction for systematic reviews of prediction modeling studies
CS: cesarean section
CTG: cardiotocogram
EHG: electrohysterogram
EPV: events per variable
IRR: incidence rate ratio
LR: logistic regression
MeSH: Medical Subject Heading
MLP-BIOM: guidelines for developing and reporting machine learning predictive models in biomedical research
OR: odds ratio
PI: prediction interval

PICOTS: population, index, comparator, outcomes, timing, and setting
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: prediction model risk of bias assessment tool
ROB: risk of bias
ROC: receiver operating characteristic
TRIPOD: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

Edited by G Eysenbach; submitted 07.10.19; peer-reviewed by E Christodoulou, R Ho, WD Dotson, L Kriston; comments to author 13.03.20; revised version received 22.06.20; accepted 24.10.20; published 17.11.20.

Please cite as:
Sufriyana H, Husnayain A, Chen YL, Kuo CY, Singh O, Yeh TY, Wu YW, Su ECY
Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis
JMIR Med Inform 2020;8(11):e16503
URL: http://medinform.jmir.org/2020/11/e16503/
doi:10.2196/16503
PMID:33200995

Original Paper

Deep Learning Methodology for Differentiating Glioma Recurrence From Radiation Necrosis Using Multimodal Magnetic Resonance Imaging: Algorithm Development and Validation

Yang Gao[1*], PhD; Xiong Xiao[2*], MD; Bangcheng Han[3], PhD; Guilin Li[4], MD; Xiaolin Ning[3], PhD; Defeng Wang[3], PhD; Weidong Cai[5], PhD; Ron Kikinis[6,7,8], MD, PhD; Shlomo Berkovsky[9], PhD; Antonio Di Ieva[10], MD, PhD; Liwei Zhang[2], MD; Nan Ji[2], MD; Sidong Liu[9], PhD

[1]Beijing Academy of Quantum Information Sciences, Beijing, China

[2]Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China

[3]School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing, China

[4]Department of Neuropathology, Beijing Neurosurgical Institute, Capital Medical University, Beijing, China

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

[6]Surgical Planning Laboratory, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

[7]Department of Computer Science, University of Bremen, Bremen, Germany

[8]Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

[9]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

[10]Computational NeuroSurgery Lab, Department of Clinical Medicine, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, Australia

[*]these authors contributed equally

Corresponding Author:
Sidong Liu, PhD
Centre for Health Informatics, Australian Institute of Health Innovation
Macquarie University
75 Talavera Road
Macquarie Park
Sydney, 2113
Australia
Phone: 61 29852729
Email: dr.sidong.liu@gmail.com

Abstract

Background: The radiological differential diagnosis between tumor recurrence and radiation-induced necrosis (ie, pseudoprogression) is of paramount importance in the management of glioma patients.

Objective: This research aims to develop a deep learning methodology for automated differentiation of tumor recurrence from radiation necrosis based on routine magnetic resonance imaging (MRI) scans.

Methods: In this retrospective study, 146 patients who underwent radiation therapy after glioma resection and presented with suspected recurrent lesions at the follow-up MRI examination were selected for analysis. Routine MRI scans were acquired from each patient, including T1, T2, and gadolinium-contrast-enhanced T1 sequences. Of those cases, 96 (65.8%) were confirmed as glioma recurrence on postsurgical pathological examination, while 50 (34.2%) were diagnosed as necrosis. A light-weighted deep neural network (DNN) (ie, efficient radionecrosis neural network [ERN-Net]) was proposed to learn radiological features of gliomas and necrosis from MRI scans. Sensitivity, specificity, accuracy, and area under the curve (AUC) were used to evaluate performance of the model in both image-wise and subject-wise classifications. Preoperative diagnostic performance of the model was also compared to that of the state-of-the-art DNN models and five experienced neurosurgeons.

Results: DNN models based on multimodal MRI outperformed single-modal models. ERN-Net achieved the highest AUC in both image-wise (0.915) and subject-wise (0.958) classification tasks. The evaluated DNN models achieved an average sensitivity of 0.947 (SD 0.033), specificity of 0.817 (SD 0.075), and accuracy of 0.903 (SD 0.026), which were significantly better than the tested neurosurgeons (P=.02 in sensitivity and P<.001 in specificity and accuracy).

Conclusions: Deep learning offers a useful computational tool for the differential diagnosis between recurrent gliomas and necrosis. The proposed ERN-Net model, a simple and effective DNN model, achieved excellent performance on routine MRI scans and showed a high clinical applicability.

(JMIR Med Inform 2020;8(11):e19805) doi:10.2196/19805

KEYWORDS

recurrent tumor; radiation necrosis; progression; pseudoprogression; multimodal MRI; deep learning

Introduction

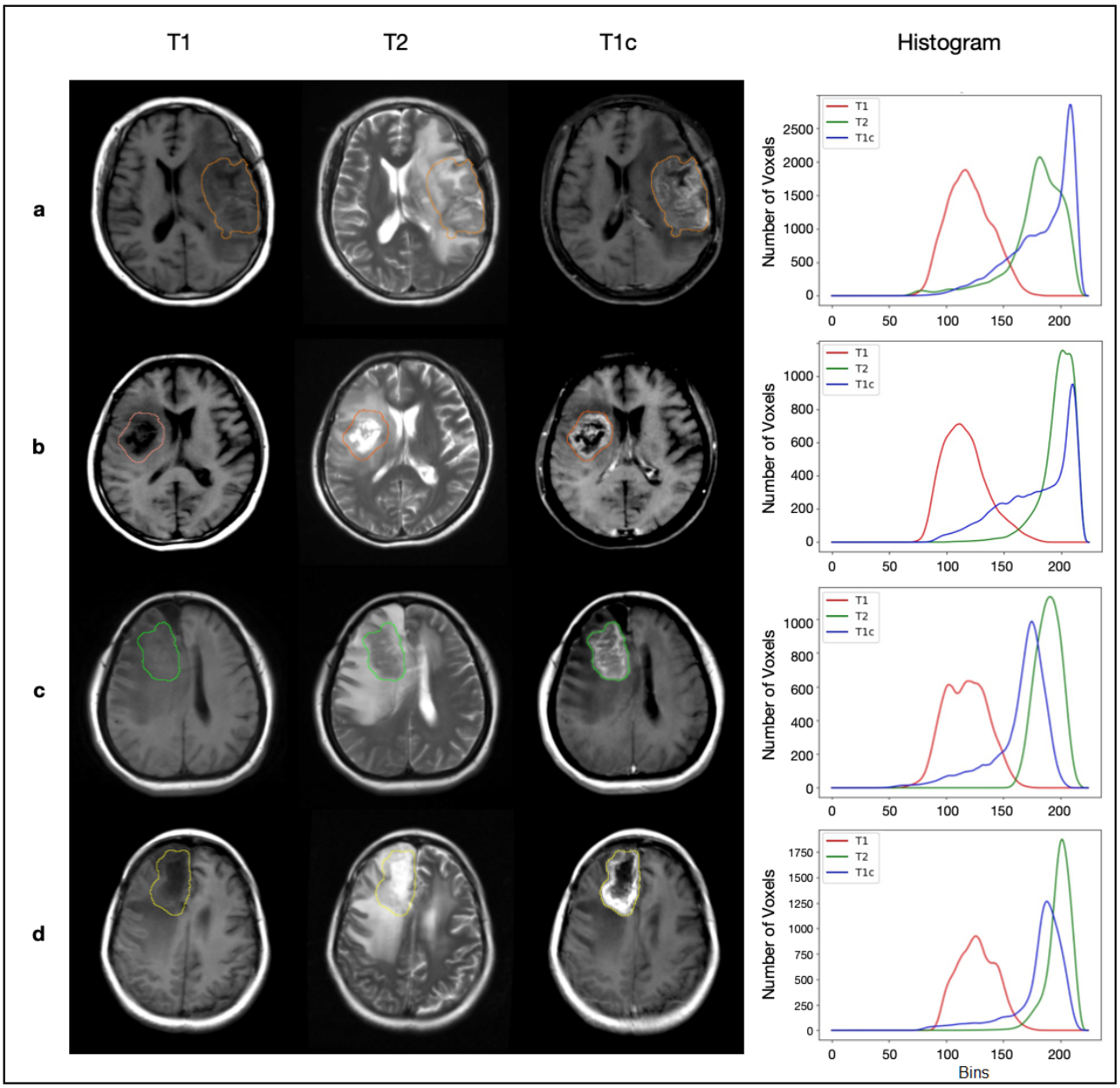
Brain radiation necrosis (ie, pseudoprogression) can be a consequence of radiation therapy, which is used for the treatment of brain tumors, with an incidence of 3%-24% [1-4]. It is of paramount importance to distinguish radiation necrosis from tumor recurrence, as these two pathologies share similar appearances in neuroimaging yet have different treatments and outcomes [5,6]. Currently, various imaging modalities, such as magnetic resonance spectroscopy (MRS) [7,8], perfusion-weighted imaging (PWI) [9], diffusion-weighted imaging (DWI) [10], and positron emission tomography (PET) with different tracers [11,12], have been applied for differentiating radiation necrosis from tumor recurrence; yet their efficacy and reliability still need further validation. Differential diagnosis between recurrent tumors and necrosis remains a major challenge in neuro-oncology and neuroradiology [1,2,5,6,13].

Recent studies demonstrate that although radiologists may not be able to systematically identify differences in the highly variable appearances of brain tumors and radionecrosis, handcrafted features extracted from routine magnetic resonance imaging (MRI) can effectively differentiate these two conditions [14-16]. As shown in these studies, handcrafted radiomic features can capture the variations in image intensity, shape, and volume and have shown promising results (see Figure 1). However, there are two major limitations that may restrict the use of these methods in the clinical setting. The first limitation is that all these methods require manual segmentation of the lesion (ie, drawing regions of interest [ROIs] of the lesion on T1-weighted MRI [T1], gadolinium-contrast-enhanced T1 [T1c], and/or T2-weighted MRI [T2]/fluid-attenuated inversion recovery [FLAIR]), from which the texture or shape features can be extracted [17]. The ROI segmentation is time-consuming and operator dependent, introducing human interference and potential noise into the analysis. Furthermore, handcrafted features extracted in these studies are usually redundant and require feature selection, which, if inaccurate, may bias the analysis.

Deep learning is a data-driven approach that uses deep neural network (DNN) models to learn the feature representations at multiple levels of abstraction [18]. Deep learning models, such as Visual Geometry Group (VGG) [19], residual neural network (ResNet) [20], and Inception [21], have substantially improved the state of the art in many visual analysis tasks (eg, ImageNet Large Scale Visual Recognition Challenge [22]), compared to handcrafted features. Deep learning methods have also demonstrated human-level performance in medical image computing, such as skin cancer classification [23], diabetic retinopathy grading [24], glaucoma detection [25], early diagnosis of Alzheimer disease [26], and, most recently, COVID-19 severity assessment [27]. Yet to the best of our knowledge, the application of deep learning in differentiating glioma recurrence from postradiotherapy necrosis has not been investigated so far.

Therefore, in this work we aim to explore the potential benefit of deep learning algorithms for distinguishing between radionecrosis and tumor recurrence using routine MRI scans. We proposed a novel DNN model (ie, efficient radionecrosis neural network [ERN-Net]) to automatically characterize the features of gliomas and necrosis from MRI images and to classify the lesions at image-based and subject-based levels, which outperformed the human experts (ie, neurosurgeons) and the state-of-the-art DNN models. Furthermore, the proposed method does not depend on lesion segmentation or any handcrafted features and, therefore, may have a higher clinical applicability.

Figure 1. The T1, T2, and T1c magnetic resonance imaging (MRI) sequences of 4 patients with their histograms of the voxels within the lesion masks. Patients (a) and (b) represent recurrent tumors; patients (c) and (d) represent radionecrosis lesions. The lesion masks were manually drawn using the software ITK-SNAP, generally used for delineating regions of interest. The histograms were created for individual sequences and further smoothed using the Hann filter. ITK: Insight Toolkit. T1: T1-weighted MRI; T1c: gadolinium-contrast-enhanced T1-weighted MRI; T2: T2-weighted MRI.



Methods

Patient Data and Imaging Protocol

This study was approved by the Institutional Review Board of Beijing Tiantan Hospital, Capital Medical University (BTH-CMU), China, and the requirement for informed consent was waived by the board as this research involves no more than minimal risk. The criteria for selecting the patient cohorts are shown in Figure 2.

We retrospectively identified patients who underwent brain tumor resection between January 2010 and November 2018, confirmed by pathology examination to be gliomas. Among the selected patients, we further selected the ones who underwent subsequent radiation therapy and presented with suspected recurrent lesions on radiological follow-up. All the patients included in this study underwent a second surgery to differentiate glioma recurrence from radiation necrosis. Histopathologic diagnoses of both the initial and recurrent lesions were performed by neuropathologists at BTH-CMU. Patients were excluded from the study if their histopathological analyses showed a mixture of tumor and necrosis.

A cohort of 146 patients were identified using our criteria. Of those, 96 (65.8%) patients were diagnosed to be affected by recurrent glioma, and 50 (34.2%) by necrosis. Of the 146 patients, 117 subjects (80.1%) were randomly assigned to the training set, and the remaining 29 subjects (19.9%) were retained as the test set. It is a common practice to split the cohort into a training set and a test set in machine learning studies, and the training set to test set ratio usually varies from 60:40 to 90:10 [20,23,26]. In this study, we chose the 80:20 split ratio to

balance the number of cases that can be used to train the model and the workload on the human experts to assess the test cases. Table 1 shows the demographic data of the subjects in the cohort as well as the distribution of the cases in the training and test data sets. The histopathological analysis results of the recurrent lesions, either recurrent tumor or necrosis, were used to categorize patients' imaging data.

Figure 2. The selection process for the patient cohorts in this study. MRI: magnetic resonance imaging; T1: T1-weighted MRI; T1c: gadolinium-contrast-enhanced T1-weighted MRI; T2: T2-weighted MRI.

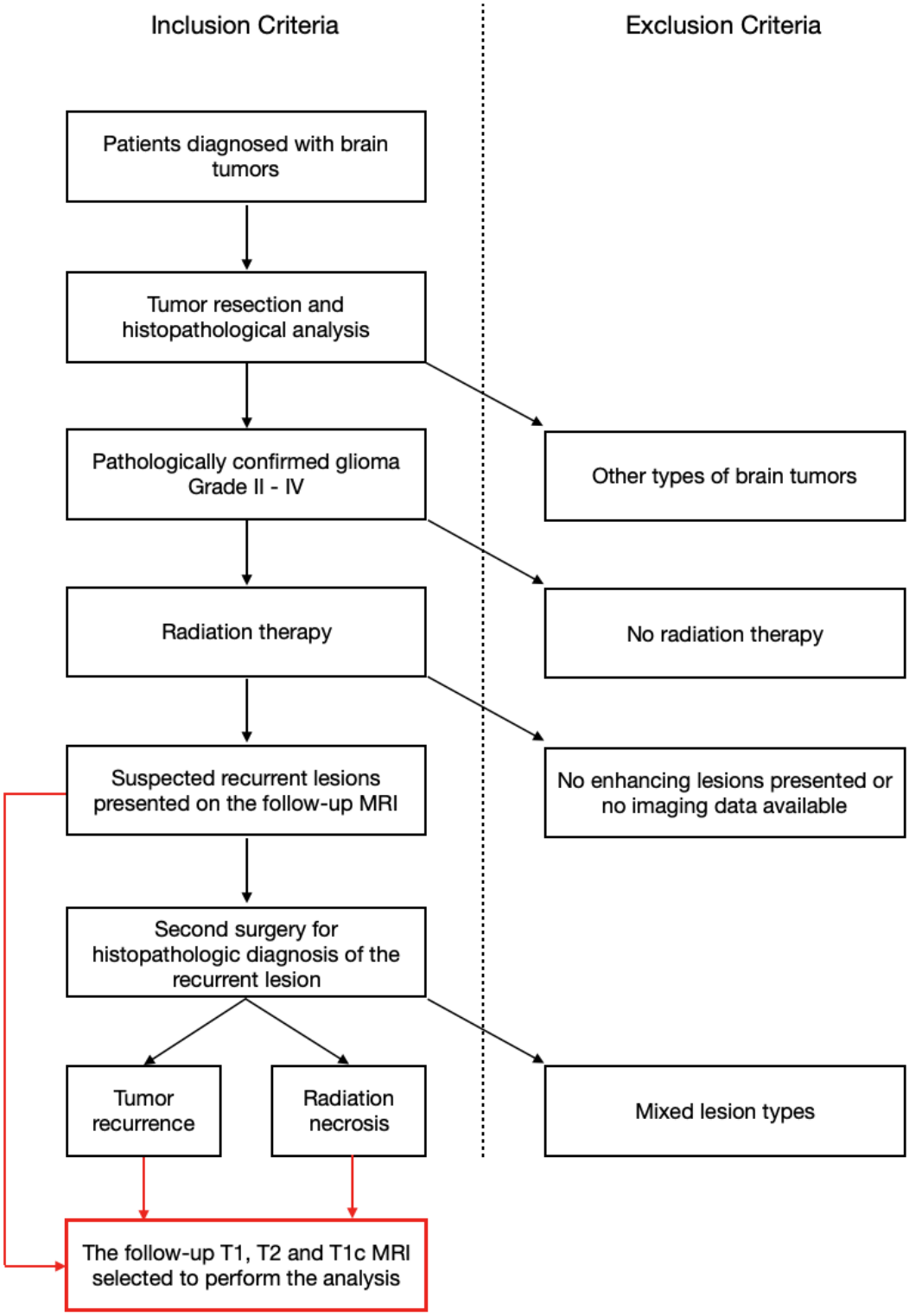


Table 1. Demographic and clinical data of the patient cohorts enrolled in this study.

Characteristic	Training set (n=117)	Test set (n=29)	Total (N=146)
Sample size (N=146), n (%)	117 (80.1)	29 (19.9)	146 (100)
Age in years, mean (SD)	40.9 (12.4)	42.0 (9.9)	41.1 (11.9)
Gender, n (%)			
Male	63 (53.8)	15 (52)	78 (53.4)
Female	54 (46.2)	14 (48)	68 (46.6)
Diagnosis of primary lesion, n (%)			
Grade II	33 (28.2)	8 (28)	41 (28.1)
Grade III	26 (22.2)	6 (21)	32 (21.9)
Grade IV	45 (38.5)	11 (38)	56 (38.4)
Unknown	13 (11.1)	4 (14)	17 (11.6)
Diagnosis of recurrent lesion, n (%)			
Necrosis	40 (34.2)	10 (34)	50 (34.2)
Glioma	77 (65.8)	19 (66)	96 (65.8)

The follow-up MRI scans of the identified patients prior to the second surgery for histopathologic diagnosis were selected to perform the analysis. The MRI data were acquired from five MRI systems at BTH-CMU. The specifications of the imaging data are listed in Table 2. All the patients have the axial T1, T2, and T1c sequences, acquired during routine clinical visits. A total of 42 MRI scans were acquired using the MAGNETOM Trio, A Tim system (Siemens), 28 scans using the MAGNETOM Verio system (Siemens), 25 scans using the Discovery MR750 system (GE Healthcare), 29 scans using the GENESIS SIGNA system (GE Healthcare) with 3 T magnetic field, and 22 using the SIGNA system (GE Healthcare) with 1.5 T magnetic field.

Table 2. Specifications of the imaging data acquired from the different magnetic resonance imaging systems.

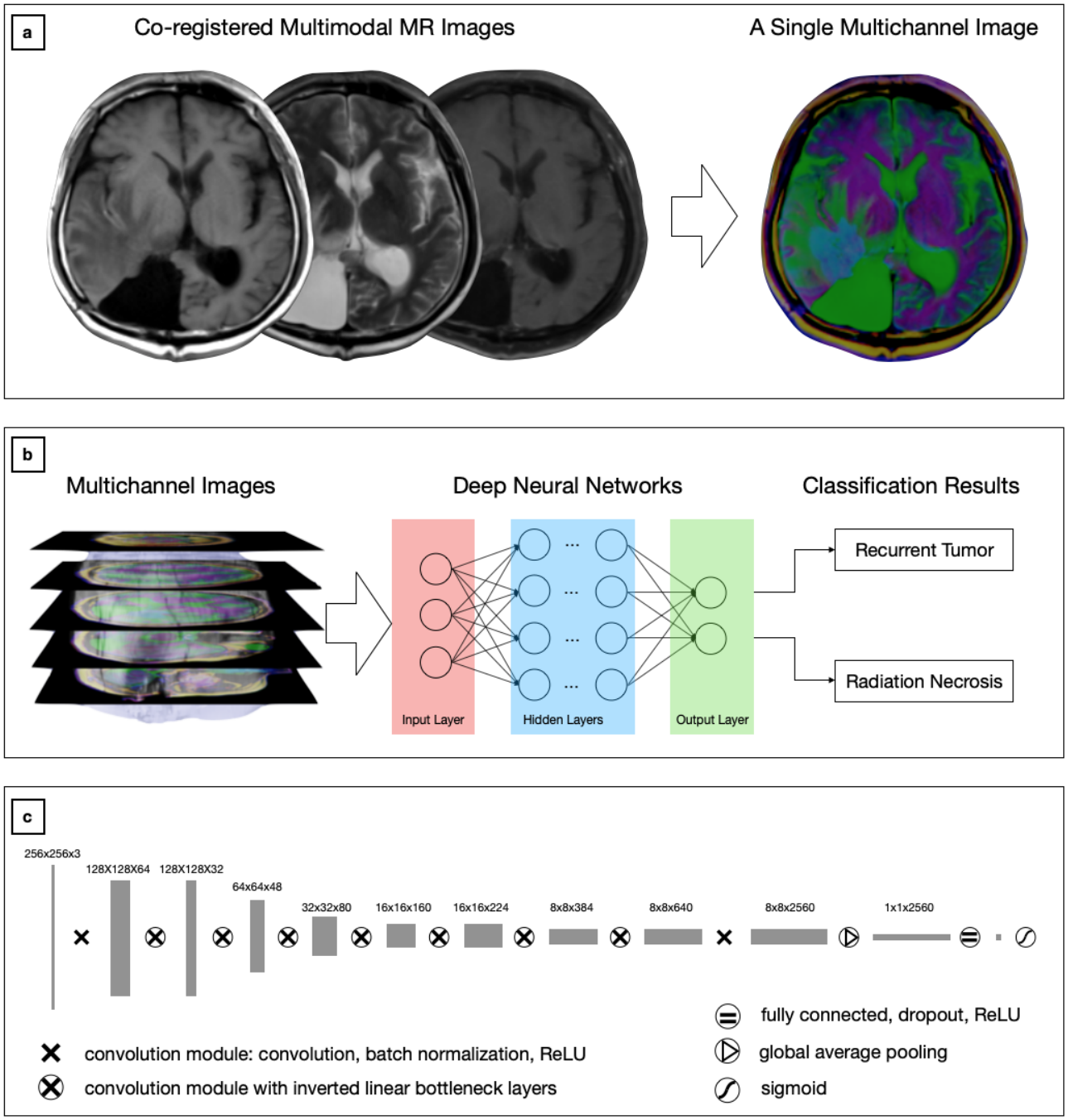
Imaging system	Field of view, mm	Slice thickness, mm	Slice spacing, mm	Matrix size
Siemens MAGNETOM Trio Tim	220	5.0	6.5	496 × 512
Siemens MAGNETOM Verio	220	5.0	6.0	496 × 512
GE Healthcare Discovery MR750	240	5.0	6.5	512 × 512
GE Healthcare GENESIS SIGNA 3 T	240	5.0	6.0	512 × 512
GE Healthcare SIGNA 1.5 T	240	5.5	6.5	512 × 512

Data Preprocessing

To standardize the MRI data across multiple MRI systems, the following preprocessing pipeline was used. First, the imaging data were corrected for bias field using the improved nonparametric, nonuniform-intensity normalization algorithm [28] built into the Advanced Normalization Tools suite of tools for brain and image analysis [29]. Second, for every patient's MRI data, the T1c and T2 images were coregistered to the T1 space using the Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library (FSL) FMRIB Linear Image Registration Tool (FLIRT) pipeline with a 6-degree-of-freedom transform [30,31]. Finally, the magnetic resonance images were linearly mapped and resampled to the Montreal Neurological Institute 152 template [32], also using FSL FLIRT, in order to make the dimensions and orientation of all the images uniform.

The MRI slices presenting enhancing lesions were identified by neuroradiologists or neurosurgeons; the multimodal magnetic resonance slices—T1, T2, and T1c—were then fused into multichannel images, as shown in Figure 3 (a). To minimize the interrater variance, we requested that the radiologists and neurosurgeons use 3D Slicer, version 4.6.2 [33], to place a marker on the axial slices if they saw a suspected recurrent lesion on the slice. Therefore, no manual outlining of the lesion was performed, taking less than two minutes for a radiologist to review an MRI image and identify the slices containing the lesion. These annotations provided by experienced neurosurgeons were used as the *ground truth* for evaluating the performance of the classifier.

Figure 3. Overview of the proposed approach. (a) The co-registered multimodal images were fused as a multichannel RGB image with T1, T2, and T1c images representing the Red, Green and Blue channels, respectively. (b) The multichannel magnetic resonance (MR) images were used to train the deep neural network (DNN) models that classified the test MR images as either a recurrent tumor or radiation necrosis. (c) Architecture of the proposed efficient radionecrosis neural network (ERN-Net). ReLU: rectified linear unit; T1: T1-weighted magnetic resonance imaging (MRI); T1c: gadolinium-contrast-enhanced T1-weighted MRI; T2: T2-weighted MRI.



The MRI slices containing the lesion were identified manually by an experienced neurosurgeon (NJ) and further reviewed by an imaging analyst (YG) based on the T1c scans. The axial T1c and corresponding T2 and T1 slices were then saved as 2D multichannel images for further analysis. A total of 5824 multichannel images, each consisting of a T1, a T2, and a T1c slice, were extracted from the 117 patients in the training set, and 1472 multichannel images were extracted from the 29 patients in the test set. The multichannel images were used to train the DNN models, which were subsequently applied to predict the patients' lesion types in the test set based on their imaging data, as shown in Figure 3 (b).

Efficient Radionecrosis Neural Network

DNN models can be considered as mathematical functions with numerous parameters. For image classification, DNN models usually use pixel values as the input features. The neurons in the hidden layers of the DNN are responsible for transforming lower-level features to higher-level features that can be used for classification. While training a DNN model, the training images and diagnostic labels (dichotomized; 0: radiation necrosis; and 1: tumor recurrence) are used to update the parameters of the model. At each training step, the model predicts the diagnostic label for an input training image, then

the prediction is compared to its ground truth label, such that the parameters of the model are modified to reduce the error on that image prediction. This process is then repeated for every image in the training set over many iterations to let the model "learn" how to differentiate the tumor recurrence signature from the necrosis one in magnetic resonance images. After the model is fully trained, it is used to infer the diagnostic probability distribution of necrosis and tumor recurrence for the test images.

For feature learning and classification, we proposed a light-weighted DNN model (ie, ERN-Net) to learn radiological features of gliomas and necrosis from MRI scans. The proposed ERN-Net model, as illustrated in Figure 3 (c), consists of only nine convolutional modules, including seven with inverted linear bottleneck layers [34]. We also benchmarked five state-of-the-art DNN models: VGG16 and VGG19 [19], ResNet-50 [20], Inception-v3 [21], and Inception-ResNet-v2 [35]. It is noteworthy that ERN-Net is 3 times smaller and 8.1 times faster than Inception-v3 [36]. All the DNN models were implemented using the TensorFlow framework, version 1.14 [37], with the ImageNet pretrained weights imported from the Keras library [38]. To address the imbalanced sample distribution, we assigned different weights to the classes during the training phase based on the ratio between the number of samples in each class and the total number of samples scaled by the number of classes (necrosis: 1.5; recurrence: 0.75). A more detailed description of these DNN models can be found in Multimedia Appendix 1.

Performance Evaluation

To evaluate the performance of the DNN models on image-wise classification, we designed an experiment in which we trained and tested these DNN models on the same data set, including 5824 training images and 1472 test images. This experiment was carried out on a per-image basis, with each image treated as an individual input sample. We also compared the performance of single-modal and multimodal MRI in the image-wise classification task. Sensitivity, specificity, accuracy, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve were used to evaluate the classification performance.

To evaluate the performance of the DNN models on a subject basis, we designed another experiment in which we aggregated the image-wise classification results to infer each subject's diagnosis. For each subject in the test set, the models that had been trained for image-wise classification in the previous experiment were reused to classify the stack of the subject's images; the image-wise classification results were then averaged as the output prediction of that subject. Performances of these DNN models in subject-wise classification were also compared with those of the human experts.

Results

Image-wise Classification

Table 3 shows the summary of the comparison of different MRI sequences using DNN models. T1c was the best performing sequence among the three routine MRI sequences, with consistently higher accuracy and AUC than T1 and T2 sequences across all the DNN models. T1c also achieved the highest sensitivity with VGG16 and Inception-v3 models (0.874 and 0.769, respectively), and the highest specificity with VGG19 and ResNet-50 models (both equal to 0.653). Considering AUC as a single metric that combines sensitivity and specificity, T2 performed slightly better than T1, although there was disagreement in other evaluation metrics. ERN-Net outperformed the VGG models in AUC on T1c (0.807, 95% CI 0.782-0.832), while Inception-ResNet-v2 achieved the highest AUC (0.841, 95% CI 0.818-0.864). We found that the sensitivity was higher than specificity in most models and sequences. This can be partially explained by the imbalanced sample distribution in the two classes, which might bias the models and, hence, the classification results.

Table 4 shows the performance comparison of the DNN models on multimodal MRI images. ERN-NET had the highest AUC (0.915, 95% CI 0.895-0.932), which was slightly better than Inception-ResNet-v2 (0.913, 95% CI 0.895-0.931) and substantially better than the other DNN models. Inception-ResNet-v2 achieved the highest score in sensitivity (0.925, 95% CI 0.907-0.941) and accuracy (0.867, 95% CI 0.848-0.884), while VGG16 had the highest specificity (0.826, 95% CI 0.791-0.858). The DNN models based on multimodal MRI outperformed the models based on individual MRI sequences in all the evaluation metrics. We again noticed that the sensitivity was higher than specificity for all the DNN models, with differences ranging from 0.032 (VGG16 sensitivity: 0.858; specificity: 0.826) to 0.236 (ResNet-50 sensitivity: 0.899; specificity: 0.663).

Table 3. Performance of the deep neural network (DNN) models on individual magnetic resonance imaging (MRI) sequences: T1-weighted MRI (T1), T2-weighted MRI (T2), and gadolinium-contrast-enhanced T1-weighted MRI (T1c).

DNN model and magnetic resonance sequence	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	Area under the curve (95% CI)
VGG[a]16				
T1	0.725 (0.696-0.753)	0.606 (0.562-0.648)	0.684 (0.660-0.708)	0.718 (0.689-0.747)
T2	0.690 (0.660-0.719)	0.686 (0.644-0.727)	0.689 (0.665-0.713)	0.767 (0.740-0.794)
T1c	0.874 (0.851-0.894)	0.540 (0.496-0.585)	0.759 (0.736-0.781)	0.770 (0.743-0.797)
VGG19				
T1	0.804 (0.778-0.829)	0.448 (0.404-0.492)	0.681 (0.657-0.705)	0.692 (0.663-0.721)
T2	0.743 (0.714-0.770)	0.554 (0.510-0.598)	0.678 (0.653-0.702)	0.741 (0.713-0.769)
T1c	0.800 (0.773-0.825)	0.653 (0.610-0.694)	0.749 (0.726-0.771)	0.795 (0.769-0.821)
ResNet[b]-50				
T1	0.782 (0.755-0.808)	0.584 (0.540-0.627)	0.714 (0.690-0.737)	0.732 (0.704-0.760)
T2	0.833 (0.808-0.852)	0.525 (0.480-0.569)	0.727 (0.703-0.750)	0.762 (0.735-0.789)
T1c	0.825 (0.799-0.848)	0.653 (0.610-0.694)	0.766 (0.743-0.787)	0.824 (0.800-0.848)
Inception-v3				
T1	0.724 (0.695-0.752)	0.596 (0.552-0.639)	0.680 (0.656-0.704)	0.706 (0.677-0.735)
T2	0.634 (0.603-0.665)	0.734 (0.693-0.772)	0.668 (0.644-0.693)	0.734 (0.706-0.762)
T1c	0.769 (0.741-0.795)	0.732 (0.691-0.770)	0.756 (0.733-0.778)	0.831 (0.807-0.855)
Inception-ResNet-v2				
T1	0.774 (0.746-0.800)	0.590 (0.546-0.633)	0.711 (0.687-0.734)	0.748 (0.720-0.776)
T2	0.829 (0.804-0.852)	0.529 (0.484-0.573)	0.726 (0.702-0.748)	0.804 (0.779-0.829)
T1c	0.812 (0.786-0.837)	0.722 (0.681-0.761)	0.781 (0.759-0.802)	0.841 (0.818-0.864)
ERN-Net[c]				
T1	0.704 (0.674-0.732)	0.519 (0.474-0.563)	0.640 (0.615-0.665)	0.646 (0.615-0.676)
T2	0.634 (0.603-0.665)	0.606 (0.562-0.648)	0.624 (0.599-0.649)	0.675 (0.645-0.705)
T1c	0.803 (0.777-0.828)	0.643 (0.600-0.685)	0.748 (0.725-0.770)	0.807 (0.782-0.832)

[a]VGG: Visual Geometry Group.

[b]ResNet: residual neural network.

[c]ERN-Net: efficient radionecrosis neural network.

Table 4. Performance of different deep neural network (DNN) models on the T1[a]-T2[b]-T1c[c]-fused images for image-based classification.

DNN models	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	Area under the curve (95% CI)
VGG[d]16	0.858 (0.834-0.880)	0.826 (0.791-0.858)	0.847 (0.828-0.865)	0.864 (0.842-0.886)
VGG19	0.852 (0.828-0.874)	0.704 (0.662-0.744)	0.801 (0.780-0.821)	0.828 (0.804-0.852)
ResNet[e]-50	0.899 (0.879-0.918)	0.663 (0.620-0.704)	0.818 (0.797-0.837)	0.866 (0.844-0.888)
Inception-v3	0.844 (0.819-0.866)	0.716 (0.675-0.755)	0.800 (0.778-0.820)	0.845 (0.822-0.868)
Inception-ResNet-v2	0.925 (0.907-0.941)	0.755 (0.716-0.792)	0.867 (0.848-0.884)	0.913 (0.895-0.931)
ERN-Net[f]	0.820 (0.794-0.844)	0.789 (0.751-0.824)	0.809 (0.788-0.829)	0.915 (0.895-0.932)

[a]T1: T1-weighted magnetic resonance imaging (MRI).

[b]T2: T2-weighted MRI.

[c]T1c: gadolinium-contrast-enhanced T1-weighted MRI.

[d]VGG: Visual Geometry Group.

[e]ResNet: residual neural network.

[f]ERN-Net: efficient radionecrosis neural network.

Subject-wise Classification

Table 5 shows the performance of different DNN models in the subject-wise classification task. Each of the 29 test subjects was considered as a single sample to be classified. In this experiment, the classification results of the images extracted from the same patient were averaged as the final output prediction of the subject. When the DNN models were evaluated on a per-subject basis by aggregating the subject's image stack, the performance was further improved to an average sensitivity of 0.947 (SD 0.033), specificity of 0.817 (SD 0.075), accuracy of 0.903 (SD 0.026), and AUC of 0.938 (SD 0.022). Both ERN-Net and Inception-ResNet-v2 achieved the highest AUC of 0.958. While Inception-ResNet-v2 also had higher sensitivity and accuracy, ERN-Net had higher specificity. In particular, Inception-ResNet-v2 achieved a sensitivity of 100%, indicating that all recurrent tumors identified by Inception-ResNet-v2 were correct. VGG16 tied for the highest specificity (0.900) with ERN-Net and the highest accuracy (0.931) with Inception-ResNet-v2. The DNN models had higher sensitivity than specificity, except ERN-Net, implying that ERN-Net was less affected by the imbalanced distribution of necrosis and recurrent tumor samples on the subject level.

We also compared the performance of the DNN models to that of five neurosurgeons, with 7-26 years of experience, who were presented with the same multimodal MRI scans as used to test the DNN models. The neurosurgeons were not shown the pathological analysis reports and were requested to make diagnoses based on the MRI data alone. The neurosurgeons achieved an average sensitivity of 0.768 (SD 0.109), specificity of 0.360 (SD 0.089), and accuracy of 0.628 (SD 0.075), which were significantly worse than the DNN models when measured using *t* tests (P=.02 in sensitivity and P<.001 in specificity and accuracy).

Figure 4 further shows the ROC curves and the AUC scores of the DNN models in the image-wise and subject-wise classification tasks. The red dots in Figure 4 (b) represent the neurosurgeons' sensitivity and specificity scores.

Table 5. Performance of different deep neural network (DNN) models for subject-based classification; the T1[a]-T2[b]-T1c[c]-fused images were used as the input to the models.

DNN models	Sensitivity	Specificity	Accuracy	Area under the curve
VGG[d]16	0.947	0.9	0.931	0.911
VGG19	0.947	0.8	0.897	0.911
ResNet[e]-50	0.947	0.7	0.862	0.937
Inception-v3	0.947	0.8	0.897	0.953
Inception-ResNet-v2	1.000	0.8	0.931	0.958
ERN-Net[f]	0.895	0.9	0.897	0.958
All DNNs, mean (SD)	0.947 (0.033)	0.817 (0.075)	0.903 (0.026)	0.938 (0.022)
All neurosurgeons, mean (SD)	0.768 (0.109)	0.360 (0.089)	0.628 (0.750)	N/A[g]
P values for *t* tests between the DNNs and the neurosurgeons	.02	<.001	<.001	N/A

[a]T1: T1-weighted magnetic resonance imaging (MRI).

[b]T2: T2-weighted MRI.

[c]T1c: gadolinium-contrast-enhanced T1-weighted MRI.

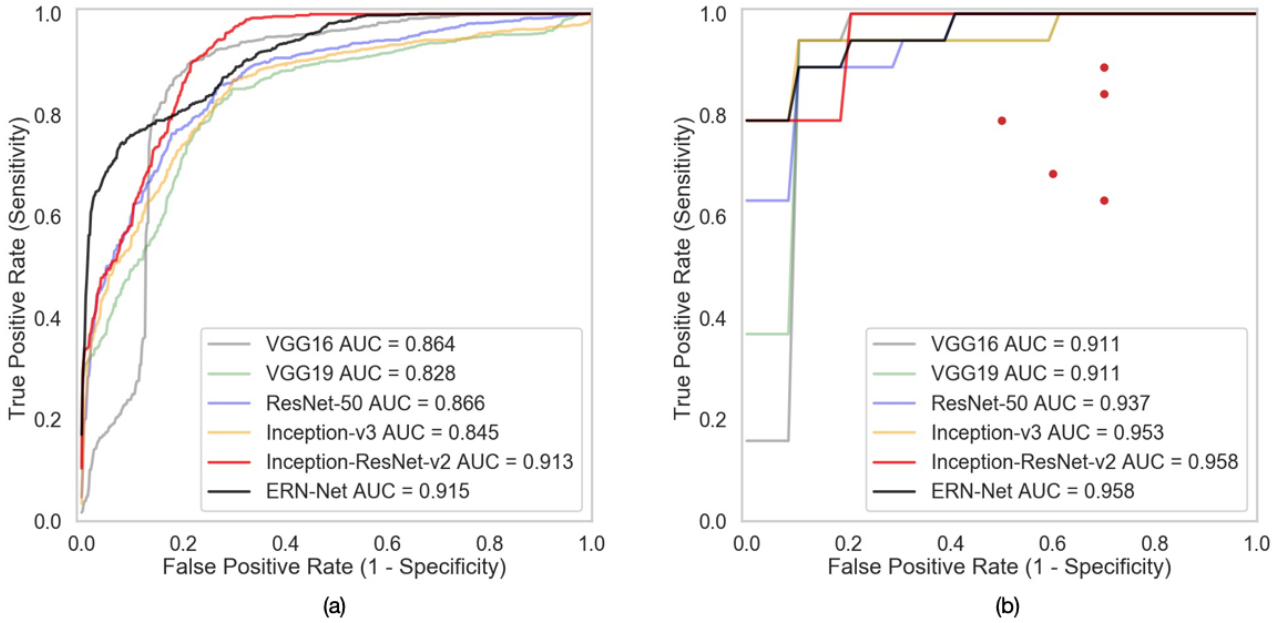
[d]VGG: Visual Geometry Group.

[e]ResNet: residual neural network.

[f]ERN-Net: efficient radionecrosis neural network.

[g]N/A: not applicable. The diagnoses made by neurosurgeons are definite (ie, yes or no), unlike those made by the DNN models (eg, 30% yes or 70% no); therefore, the area under the curve cannot be computed without a probability distribution of predictions.

Figure 4. Plots showing (a) performance of the deep neural network (DNN) models on multimodal magnetic resonance imaging in the image-based classification task and (b) performance of the DNN models and neurosurgeons in the subject-based classification task. Performance of the DNN models was evaluated using the area under the curve (AUC) of the receiver operating characteristic curves, while the five neurosurgeons' sensitivity and specificity scores are represented by the red dots. ERN-Net: efficient radionecrosis neural network; ResNet: residual neural network; VGG: Visual Geometry Group.



Discussion

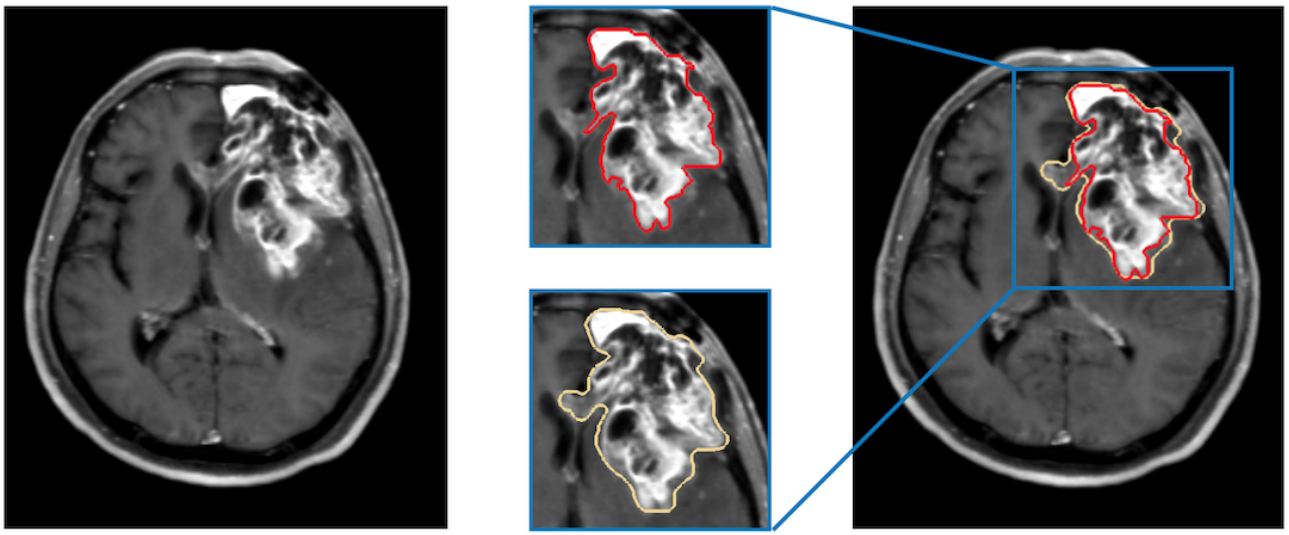
Principal Findings

To the best of our knowledge, this is the first research on the application of DNN models to routine MRI scans for the purposes of automated differentiation between radiation necrosis and recurrent tumors. We found that T1c is the most informative routine MRI sequence for identifying radiation necrosis, which aligns well with many previous studies [1,2,5,6,13,15]. However, other routine MRI sequences, including T1 and T2, also provide useful and complementary information to T1c in characterizing the tumors and necrosis, as evidenced by the improved performance of the combined MRI sequences.

The proposed ERN-Net model achieved the highest AUC in both image-wise classification (0.915) and subject-wise classification (0.958), while being substantially smaller and faster compared to the other DNN models. Overall, the DNN models achieved better performance than the human experts. The most important advantage of the DNN models is that they have a higher discriminative power in recognizing radiation necrosis with a mean specificity of 0.817 (SD 0.075) compared to the mean specificity of 0.360 (SD 0.089) achieved by experienced neurosurgeons (P<.001).

Compared to previously reported machine learning methods, which were generally based on handcrafted features and user-defined classifiers [14-16], DNN models use an end-to-end approach to integrate feature learning and classification and, therefore, could eliminate the dependence on the selected feature descriptors and classifiers. Furthermore, the proposed method does not require manual drawing of the lesion, which is time-consuming and may result in interreader variance [17,39], as shown in Figure 5. We proposed a lesional slice identification approach to select the relevant slices instead of creating the lesion masks manually. This approach reduced the time required for annotating tumor masks and can also capture contextual spatial information of the perilesional tissues. Both the trained DNN models and the lesion slice identification module support cross-platform systems and can be seamlessly integrated into existing image analysis and reporting workstations within a hospital, aiming to generate differential diagnosis reports automatically. More importantly, the performance of the proposed method is substantially higher than that of the previously reported methods [14-16].

Figure 5. A T1c tumor image and its corresponding tumor masks created by two neuroradiologists independently, which shows the disagreement between annotators. MRI: magnetic resonance imaging; T1c: gadolinium-contrast-enhanced T1-weighted MRI.



Currently, there exist other imaging techniques for differential diagnosis of recurrent tumor and radiation necrosis, such as MRS [7,8], PWI [9], DWI [10], and PET [11,12], yet none of them demonstrate sufficiently high efficacy for clinical use. A meta-analysis on PET showed that L-[*methyl*-^{11}C]methionine (11C-MET) PET achieved promising results, with a pooled sensitivity and specificity of 0.880 (95% CI 0.850-0.910) and 0.850 (95% CI 0.800-0.890), respectively, and a summary receiver operating characteristic (SROC) score of 0.935 [12]. Another meta-analysis of 11C-MET PET showed an SROC score of 0.8914 [40]. Both PET meta-analysis studies showed a lower performance than the proposed method. In addition, the relative accessibility, radiation exposure, and higher cost of PET limit its clinical applicability. MRS demonstrated moderate diagnostic performance in differentiating glioma recurrence from radiation necrosis based on metabolite ratios, such as choline to creatinine and choline to N-acetylaspartate, and it is strongly recommended to combine MRS with other imaging technologies to improve diagnostic accuracy [3]. Previous studies on machine learning and imaging techniques have two notable limitations: first, the diagnoses included in many earlier studies were not pathologically confirmed; second, the sample sizes were too small. These limitations led to inconclusive findings, such that the differential diagnosis of tumor recurrence and necrosis is still a largely unsolved clinical problem [2,12]. To the best of our knowledge, the imaging data set (N=146) used in this study represents the largest cohort in the same kind of studies and includes pathologically confirmed diagnoses as ground truth labels; therefore, it is a more reliable data set to address this problem.

Limitations and Future Work

There are also a few limitations of this study. Although we used a larger data set for the same analysis, it is still a relatively small data set compared to the generic image data sets used in the field of computer vision. This may potentially lead to overfitting or undertraining when training a DNN model. Furthermore, due to the retrospective nature of this study, the DNN models were only trained on an imbalanced data set with readily available 2D routine MRI sequences. The imbalanced distribution of samples may induce bias in the DNN model, leading to higher sensitivities but low specificities. Although we attempted to address this issue by weighting the samples during the training phase, the models still favor positive class over the negative class. It will be beneficial to extend the sample size by including

data from other centers and using data augmentation methods to further improve and validate the proposed method. Other MRI sequences, such as FLAIR, PWI, DWI, and delayed-contrast MRI, and the 3D data set may potentially improve the classification performance of the DNN models. Last but not least, also due to the retrospective nature of this study, no glioma subtypes, such as astrocytoma, oligodendroglioma, and glioblastoma; molecular genetic features, such as isocitrate dehydrogenase and alpha thalassemia/mental retardation syndrome X-linked genes; nor 1p/19q chromosome co-deletion status [41,42] were included. These aspects should be investigated in future studies.

The proposed method has high clinical potential. Distinguishing glioma recurrence from radiation necrosis remains a critical challenge in clinical neuro-oncology. Misdiagnosing radiation necrosis as tumor recurrence may result in unnecessary surgery, whereas misdiagnosing tumor recurrence as radiation necrosis will delay the treatment of tumors. Currently, the differential diagnosis of radiation necrosis and recurrent tumor relies on histopathologic analysis, which requires biopsy or open surgery to gain tissue for the analysis. This study's method proposes a sound alternative to the second surgery for the purpose of gaining tissue for histopathologic analysis, therefore avoiding invasive operations and lowering the risks to patients. In addition, up to now, there have been no clinical guidelines for preoperative diagnosis of glioma recurrence and radiation necrosis based on routine MRI sequences. Our study underlines important insights about the imaging of recurrent tumors and radiation necrosis through examining the radiological features learned by the DNN models; hence, it is likely to take an important role in formulating the guidelines for the differential diagnosis of recurrent lesions and for glioma follow-up.

Conclusions

In this work, we demonstrated that DNN models based on multimodal MRI can differentiate radionecrosis from recurrent gliomas more effectively than models based on single MRI sequences; in addition, the DNN model's performance is significantly better than that of the tested experienced clinicians on subject-wise diagnosis. Therefore, the proposed deep learning method, which does not depend on lesion segmentation or any handcrafted features, can be a useful tool for differentiating between radiation necrosis and recurrent tumors, with a high applicability potential in the clinical setting.

Acknowledgments

This study was supported by Beijing Natural Science Foundation (4191002) and the Capital Characteristic Clinical Application Project (Z181100001718196). SL acknowledges the support of an Australian National Health and Medical Research Council (NHMRC) grant: the NHMRC Early Career Fellowship (1160760). ADI received the 2019 John Mitchell Crouch Fellowship from the Royal Australasian College of Surgeons, which, along with Macquarie University cofunding, supported the opening of the Computational NeuroSurgery Lab at Macquarie University, Sydney, Australia. Moreover, he is supported by an Australian Research Council Future Fellowship (2019-2023, FI190100623). Special thanks to Fujitsu Australia Ltd for supporting the computing facilities for this research.

Authors' Contributions

The project was initially conceptualized and supervised by NJ and SL. The patient data and imaging data were acquired by YG and XX. The histopathology results were reviewed by GL, NJ, and XX. The analysis methods were by implemented by SL. The data were analyzed by YG. The research findings were interpreted by XX and NJ. All authors were involved in the design of the work. The manuscript was drafted by YG and SL, and all authors have substantively revised it. All authors have reviewed and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the deep neural networks used in this study.
[DOC File , 41 KB - medinform_v8i11e19805_app1.doc]

References

1. Verma N, Cowperthwaite MC, Burnett MG, Markey MK. Differentiating tumor recurrence from treatment necrosis: A review of neuro-oncologic imaging strategies. Neuro Oncol 2013 May;15(5):515-534 [FREE Full text] [doi: 10.1093/neuonc/nos307] [Medline: 23325863]
2. Zikou A, Sioka C, Alexiou GA, Fotopoulos A, Voulgaris S, Argyropoulou MI. Radiation necrosis, pseudoprogression, pseudoresponse, and tumor recurrence: Imaging challenges for the evaluation of treated gliomas. Contrast Media Mol Imaging 2018;2018:6828396 [FREE Full text] [doi: 10.1155/2018/6828396] [Medline: 30627060]
3. Zhang H, Ma L, Wang Q, Zheng X, Wu C, Xu B. Role of magnetic resonance spectroscopy for the differentiation of recurrent glioma from radiation necrosis: A systematic review and meta-analysis. Eur J Radiol 2014 Dec;83(12):2181-2189. [doi: 10.1016/j.ejrad.2014.09.018] [Medline: 25452098]

4. Gao L, Xu W, Li T, Zheng J, Chen G. Accuracy of 11C-choline positron emission tomography in differentiating glioma recurrence from radiation necrosis. Medicine 2018;97(29):e11556. [doi: 10.1097/md.0000000000011556]
5. Alexiou GA, Tsiouris S, Kyritsis AP, Voulgaris S, Argyropoulou MI, Fotopoulos AD. Glioma recurrence versus radiation necrosis: Accuracy of current imaging modalities. J Neurooncol 2009 Oct;95(1):1-11. [doi: 10.1007/s11060-009-9897-1] [Medline: 19381441]
6. Giglio P, Gilbert MR. Cerebral radiation necrosis. Neurologist 2003 Jul;9(4):180-188. [doi: 10.1097/01.nrl.0000080951.78533.c4] [Medline: 12864928]
7. Kumar AJ, Leeds NE, Fuller GN, Van Tassel P, Maor MH, Sawaya RE, et al. Malignant gliomas: MR imaging spectrum of radiation therapy- and chemotherapy-induced necrosis of the brain after treatment. Radiology 2000 Nov;217(2):377-384. [doi: 10.1148/radiology.217.2.r00nv36377] [Medline: 11058631]
8. Sundgren P. MR spectroscopy in radiation injury. AJNR Am J Neuroradiol 2009 Apr 15;30(8):1469-1476. [doi: 10.3174/ajnr.a1580]
9. Barajas R, Chang J, Sneed P, Segal M, McDermott M, Cha S. Distinguishing recurrent intra-axial metastatic tumor from radiation necrosis following gamma knife radiosurgery using dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging. AJNR Am J Neuroradiol 2008 Nov 20;30(2):367-372. [doi: 10.3174/ajnr.a1362]
10. Xu J, Li Y, Lian J, Dou S, Yan F, Wu H, et al. Distinction between postoperative recurrent glioma and radiation injury using MR diffusion tensor imaging. Neuroradiology 2010 Dec 23;52(12):1193-1199. [doi: 10.1007/s00234-010-0731-4] [Medline: 20571787]
11. Takenaka S, Asano Y, Shinoda J, Nomura Y, Yonezawa S, Miwa K, et al. Comparison of (11)C-methionine, (11)C-choline, and (18)F-fluorodeoxyglucose-PET for distinguishing glioma recurrence from radiation necrosis. Neurol Med Chir (Tokyo) 2014;54(4):280-289 [FREE Full text] [doi: 10.2176/nmc.oa2013-0117] [Medline: 24305028]
12. Xu W, Gao L, Shao A, Zheng J, Zhang J. The performance of 11C-methionine PET in the differential diagnosis of glioma recurrence. Oncotarget 2017 Oct 31;8(53):91030-91039 [FREE Full text] [doi: 10.18632/oncotarget.19024] [Medline: 29207622]
13. Mullins ME, Barest GD, Schaefer PW, Hochberg FH, Gonzalez RG, Lev MH. Radiation necrosis versus glioma recurrence: Conventional MR imaging clues to diagnosis. AJNR Am J Neuroradiol 2005 Sep;26(8):1967-1972 [FREE Full text] [Medline: 16155144]
14. Tiwari P, Prasanna P, Wolansky L, Pinho M, Cohen M, Nayate A, et al. Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI: A feasibility study. AJNR Am J Neuroradiol 2016 Sep 15;37(12):2231-2236. [doi: 10.3174/ajnr.a4931]
15. Zhang Z, Yang J, Ho A, Jiang W, Logan J, Wang X, et al. A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images. Eur Radiol 2018 Jun 24;28(6):2255-2263 [FREE Full text] [doi: 10.1007/s00330-017-5154-8] [Medline: 29178031]
16. Ismail M, Hill V, Statsevych V, Huang R, Prasanna P, Correa R, et al. Shape features of the lesion habitat to differentiate brain tumor progression from pseudoprogression on routine multiparametric MRI: A multisite study. AJNR Am J Neuroradiol 2018 Nov 01;39(12):2187-2193. [doi: 10.3174/ajnr.a5858]
17. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 2015 Oct;34(10):1993-2024 [FREE Full text] [doi: 10.1109/TMI.2014.2377694] [Medline: 25494501]
18. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
19. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations. 2015 May 7 Presented at: International Conference on Learning Representations; May 7-9, 2015; San Diego, CA URL: http://arxiv.org/abs/1409.1556
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.: IEEE; 2016 Jun 27 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV p. 770-778 URL: https://ieeexplore.ieee.org/document/7780459 [doi: 10.1109/CVPR.2016.90]
21. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.: IEEE; 2016 Jun 27 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV p. 2818-2826 URL: https://ieeexplore.ieee.org/document/7780677 [doi: 10.1109/cvpr.2016.308]
22. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015 Apr 11;115(3):211-252. [doi: 10.1007/s11263-015-0816-y]
23. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Jan 25;542(7639):115-118. [doi: 10.1038/nature21056]
24. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]

25. Liu S, Graham SL, Schulz A, Kalloniatis M, Zangerl B, Cai W, et al. A deep learning-based algorithm identifies glaucomatous discs using monoscopic fundus photographs. Ophthalmol Glaucoma 2018;1(1):15-22. [doi: 10.1016/j.ogla.2018.04.002] [Medline: 32672627]
26. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, ADNI. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans Biomed Eng 2015 Apr;62(4):1132-1140 [FREE Full text] [doi: 10.1109/TBME.2014.2372011] [Medline: 25423647]
27. Feng Y, Liu S, Cheng Z, Quiroz J, Rezazadegan D, Chen P, et al. Severity assessment and progression prediction of COVID-19 patients based on the LesionEncoder framework and chest CT. medRxiv. 2020 Aug 06. URL: https://www.medrxiv.org/content/10.1101/2020.08.03.20167007v2.full.pdf [accessed 2020-11-09]
28. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 bias correction. IEEE Trans Med Imaging 2010 Jun;29(6):1310-1320. [doi: 10.1109/tmi.2010.2046908]
29. Avants BB, Tustison NJ, Stauffer M, Song G, Wu B, Gee JC. The Insight ToolKit image registration framework. Front Neuroinform 2014;8:44 [FREE Full text] [doi: 10.3389/fninf.2014.00044] [Medline: 24817849]
30. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal 2001 Jun;5(2):143-156. [doi: 10.1016/s1361-8415(01)00036-6]
31. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 2002 Oct;17(2):825-841. [doi: 10.1006/nimg.2002.1132]
32. Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL, Brain Development Cooperative Group. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage 2011 Jan 01;54(1):313-327 [FREE Full text] [doi: 10.1016/j.neuroimage.2010.07.033] [Medline: 20656036]
33. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012 Nov;30(9):1323-1341 [FREE Full text] [doi: 10.1016/j.mri.2012.05.001] [Medline: 22770690]
34. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.: IEEE; 2018 Jun 18 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18-23, 2018; Salt Lake City, UT p. 4510-4520 URL: https://ieeexplore.ieee.org/document/8578572 [doi: 10.1109/cvpr.2018.00474]
35. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017 Feb 04 Presented at: Thirty-First AAAI Conference on Artificial Intelligence; February 4-9, 2017; San Francisco, CA p. 4278-4284 URL: https://dl.acm.org/doi/10.5555/3298023.3298188
36. Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the Thirty-Sixth International Conference on Machine Learning. 2019 Jun 15 Presented at: Thirty-Sixth International Conference on Machine Learning; June 9-15, 2019; Long Beach, CA p. 6105-6114 URL: http://proceedings.mlr.press/v97/tan19a.html
37. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation.: USENIX Association; 2016 Nov 2 Presented at: 12th USENIX Conference on Operating Systems Design and Implementation; November 2-4, 2016; Savannah, GA p. 265-283 URL: https://dl.acm.org/doi/10.5555/3026877.3026899
38. Keras. URL: https://keras.io/ [accessed 2020-11-09]
39. Russo C, Liu S, Di Ieva A. Spherical coordinates transformation pre-processing in deep convolution neural networks for brain tumor segmentation in MRI. arXiv. 2020 Aug 17. URL: https://arxiv.org/pdf/2008.07090 [accessed 2020-11-09]
40. Wang X, Hu X, Xie P, Li W, Li X, Ma L. Comparison of magnetic resonance spectroscopy and positron emission tomography in detection of tumor recurrence in posttreatment of glioma: A diagnostic meta-analysis. Asia Pac J Clin Oncol 2015 Jun;11(2):97-105. [doi: 10.1111/ajco.12202] [Medline: 24783970]
41. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A summary. Acta Neuropathol 2016 Jun;131(6):803-820. [doi: 10.1007/s00401-016-1545-1] [Medline: 27157931]
42. Liu S, Shah Z, Sav A, Russo C, Berkovsky S, Qian Y, et al. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci Rep 2020 May 07;10(1):7733. [doi: 10.1038/s41598-020-64588-y] [Medline: 32382048]

Abbreviations

11C-MET: L-[*methyl*-^{11}C]methionine
AUC: area under the curve
BTH-CMU: Beijing Tiantan Hospital, Capital Medical University
DNN: deep neural network
DWI: diffusion-weighted imaging
ERN-Net: efficient radionecrosis neural network

FLAIR: fluid-attenuated inversion recovery
FLIRT: Functional Magnetic Resonance Imaging of the Brain Linear Image Registration Tool
FMRIB: Functional Magnetic Resonance Imaging of the Brain
FSL: Functional Magnetic Resonance Imaging of the Brain Software Library
MRI: magnetic resonance imaging
MRS: magnetic resonance spectroscopy
NHMRC: National Health and Medical Research Council
PET: positron emission tomography
ResNet: residual neural network
ROC: receiver operating characteristic
ROI: region of interest
SROC: summary receiver operating characteristic
T1: T1-weighted magnetic resonance imaging
T1c: gadolinium-contrast-enhanced T1-weighted magnetic resonance imaging
T2: T2-weighted magnetic resonance imaging
VGG: Visual Geometry Group

Edited by C Lovis; submitted 02.05.20; peer-reviewed by M Feng, R Dewey; comments to author 07.06.20; revised version received 31.08.20; accepted 27.09.20; published 17.11.20.

Please cite as:
Gao Y, Xiao X, Han B, Li G, Ning X, Wang D, Cai W, Kikinis R, Berkovsky S, Di Ieva A, Zhang L, Ji N, Liu S
Deep Learning Methodology for Differentiating Glioma Recurrence From Radiation Necrosis Using Multimodal Magnetic Resonance Imaging: Algorithm Development and Validation
JMIR Med Inform 2020;8(11):e19805
URL: http://medinform.jmir.org/2020/11/e19805/
doi:10.2196/19805
PMID:33200991

Original Paper

Development of an Artificial Intelligence–Based Automated Recommendation System for Clinical Laboratory Tests: Retrospective Analysis of the National Health Insurance Database

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

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

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

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

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

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

Corresponding Author:
Yu-Chuan Jack Li, MD, PhD
Graduate Institute of Biomedical Informatics, College of Medical Science and Technology
Taipei Medical University
250 Wu-Hsing St., Taipei 110
Taipei
Taiwan
Phone: 886 2 27361661 ext 7600
Email: jaak88@gmail.com

Abstract

Background: Laboratory tests are considered an essential part of patient safety as patients' screening, diagnosis, and follow-up are solely based on laboratory tests. Diagnosis of patients could be wrong, missed, or delayed if laboratory tests are performed erroneously. However, recognizing the value of correct laboratory test ordering remains underestimated by policymakers and clinicians. Nowadays, artificial intelligence methods such as machine learning and deep learning (DL) have been extensively used as powerful tools for pattern recognition in large data sets. Therefore, developing an automated laboratory test recommendation tool using available data from electronic health records (EHRs) could support current clinical practice.

Objective: The objective of this study was to develop an artificial intelligence–based automated model that can provide laboratory tests recommendation based on simple variables available in EHRs.

Methods: A retrospective analysis of the National Health Insurance database between January 1, 2013, and December 31, 2013, was performed. We reviewed the record of all patients who visited the cardiology department at least once and were prescribed laboratory tests. The data set was split into training and testing sets (80:20) to develop the DL model. In the internal validation, 25% of data were randomly selected from the training set to evaluate the performance of this model.

Results: We used the area under the receiver operating characteristic curve, precision, recall, and hamming loss as comparative measures. A total of 129,938 prescriptions were used in our model. The DL-based automated recommendation system for laboratory tests achieved a significantly higher area under the receiver operating characteristic curve (AUROCmacro and AUROCmicro of 0.76 and 0.87, respectively). Using a low cutoff, the model identified appropriate laboratory tests with 99% sensitivity.

Conclusions: The developed artificial intelligence model based on DL exhibited good discriminative capability for predicting laboratory tests using routinely collected EHR data. Utilization of DL approaches can facilitate optimal laboratory test selection for patients, which may in turn improve patient safety. However, future study is recommended to assess the cost-effectiveness for implementing this model in real-world clinical settings.

(JMIR Med Inform 2020;8(11):e24163) doi:10.2196/24163

KEYWORDS

artificial intelligence; deep learning; clinical decision-support system; laboratory test; patient safety

Introduction

Laboratory tests are key components of the health care system and patient safety [1]. These tests assist physicians, helping them make many important decisions related to prevention, diagnosis, treatment, and management of chronic diseases [2]. However, in recent years, laboratory error rates have increased significantly, which has raised serious concerns about patient's safety. Compared with other types of medical errors, laboratory errors have received little attention, despite these errors often causing significant harm to the patients [3]. Previous studies have reported that indiscriminate and inappropriate use of laboratory tests puts a significant and unnecessary burden on the health care system [4,5]. The value and associated cost of such inappropriate tests in the diagnostic and management process thus need to be determined.

Inappropriate testing can be in several forms. The first one is overutilization or overordering, which refers to recommended tests to the patients that are ordered without any indication. The second one is underutilization, which refers to recommended laboratory tests that are indicated but not ordered. Overutilization can result in unnecessary blood draws and other sample-collection procedures [6,7]. It increases the likelihood of false-positive results, which can lead to incorrect diagnoses, increased costs, and potential harm due to unwarranted additional intervention [8]. By contrast, underutilization can result in morbidity due to delayed or missed diagnoses and in downstream overutilization. Both overutilization and underutilization can lead to longer hospital stays and contribute to legal liability.

Deep learning (DL), a subset of machine learning, is being used in many areas including health care and has already shown its promise in various domains. This success can be attributed to an increase in computational power and the availability of massive amounts of data sets [9,10]. The field of DL has achieved immense success in training the machines to understand and manipulate data, including images [11], language [12], and speech [13]. In particular, health care and medicine are reaping significant benefits from the field because of the sheer volume of data generated every day in different forms. A quick and accurate laboratory test is crucial for patient's safety through successful diagnosis and proper treatment of diseases. Because DL algorithms can easily handle hundreds of thousands of attributes and are capable of detecting and utilizing their interaction, developing an automated recommendation tool is always appreciable to improve proper clinical decisions. Accordingly, our study developed and evaluated a DL algorithm–based automated recommendation system using variables available in electronic health records (EHRs). We hypothesized that the DL algorithm can capture high-dimensional, nonlinear relationships among clinical features and a laboratory test recommendation system can be developed that can help physicians prescribe laboratory tests to individual patients more accurately as well as ensure safety of these patients.

Methods

Data Sources

We collected data from the Taiwanese National Health Insurance Research and Development (NHIRD) database, which contains all claims for the medications and diagnoses data of 23 million (covers approximately 99.9% of the total population in Taiwan) Taiwanese. The database includes patients' demographic information, number of prescriptions, the brand and generic name of the drugs, the date of prescriptions, dosage of medication, and diagnosis. The quality and completeness of this database are excellent and have been used to conduct high-quality research [14-16]. This study was approved by the Taipei Medical University Research Ethical Board. Participants' consent was not required because all the individual information was deidentified.

Study Population

In this study, we retrieved prescription information for those patients who visited the cardiology department at least once from 2 million randomly selected patient's data from the NHIRD database between January 1, 2013, and December 31, 2013.

Variables Collection and Data Cleaning

We collected EHR data available at the time of ordering laboratory tests to develop the predictive model; these data included patients' demographics, visit date, department ID, diagnosis, medications, and laboratory tests. We considered the first 3 digits in ICD-9-CM to retrieve information about comorbidities. The ICD-9-CM is usually distributed from 001 to 999 and V01 to V82. Furthermore, we considered the first 5 characters of the ATC code that cover almost every medication in a single category. For example, the 5 digits ATC of the code C09AA (ACE inhibitors, plain) include all plain ACE inhibitors such as C09AA01 (captopril), C09AA02 (enalapril). However, 7 characters (e.g., R06AX12) were considered for other drugs with "X" as the fifth character because usually "X" means other agents in the ATC code. The overall data set retrieved included 328 types of laboratory tests. This is a large amount of data and most laboratory tests were not ordered frequently, which can make prediction performance worse. We therefore calculated the percentage of all laboratory tests and selected a threshold of 0.5% to be included in this study. Finally, we narrowed the laboratory tests down to 35, which contributed to at least 0.5% of all tests in the study period. However, these 35 tests contributed to more than 90% of total tests (see Table S2 in Multimedia Appendix 1). All extracted data were used to make the matrices for data normalization (see Table S1 in Multimedia Appendix 1), and then used to train the deep neural network (DNN)–based multilabel prediction model to correctly identify laboratory items.

Model Development and Validation

In this study, 80% of data were assigned to the training set, and 20% to the testing set. In the internal validation, we randomly selected 25% of the data from the training set and evaluated model performance (Figure 1). We developed the DNN model on the training set using all variables and assessed the model using the validation set to predict laboratory tests (see Figures

S1 and S2 in Multimedia Appendix 1). DNN is an algorithm in which an artificial neural network consists of multiple layers between the input layer and the output layer. The input of DNN moves through the layers calculating the probability of each output (Figure 2). We used 3 hidden layers. Activation functions used in this model were ReLU and Softmax. We used 20 epochs in our model.

Figure 1. Overall study design.

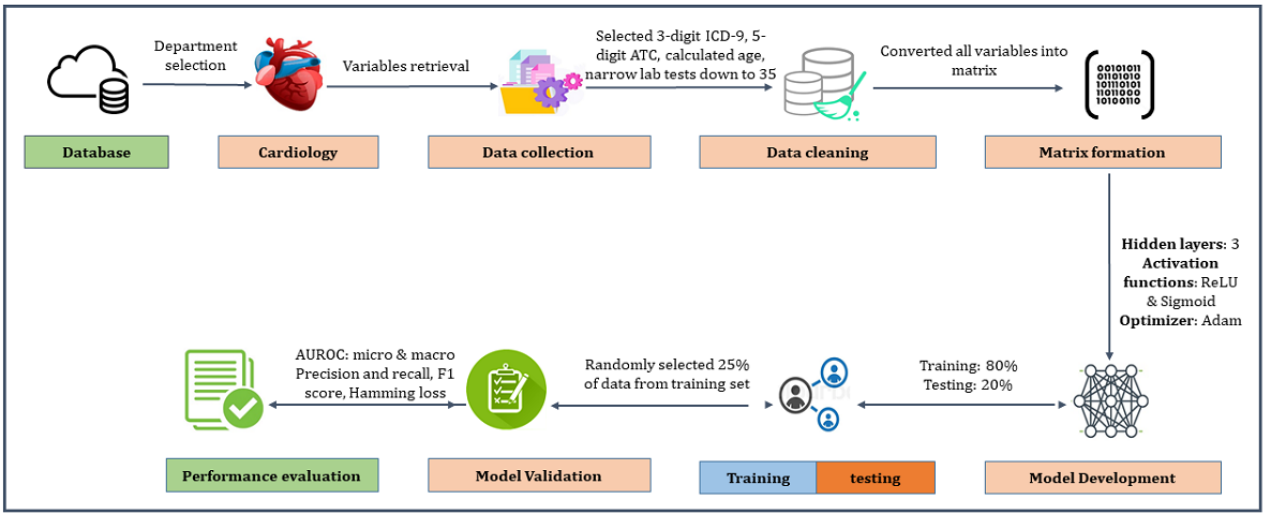
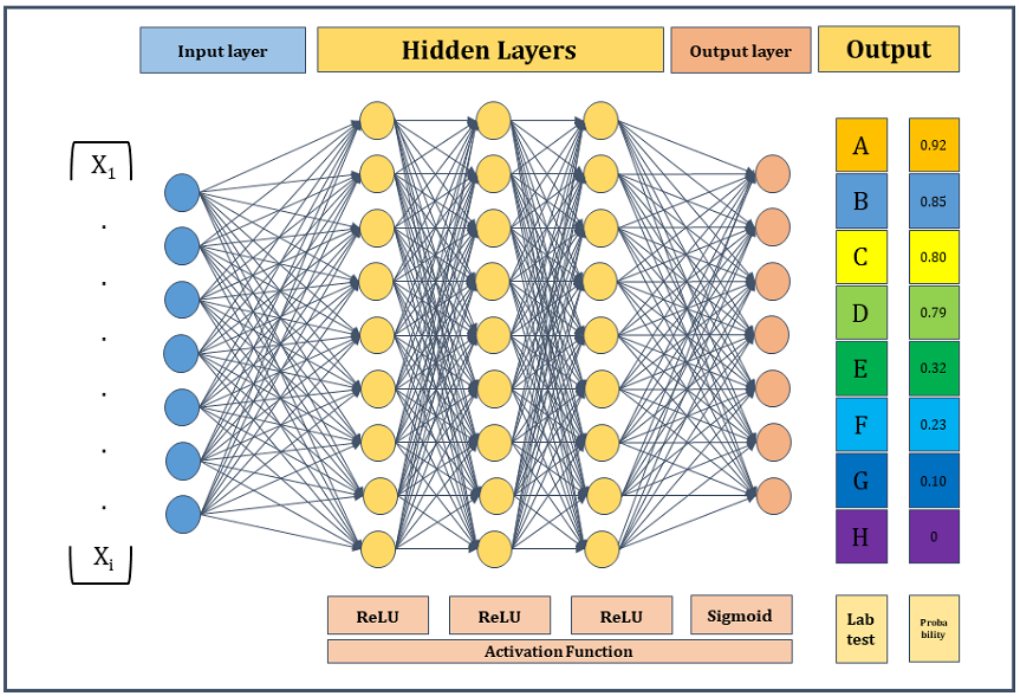


Figure 2. An architecture of proposed deep learning model.



Activation Function

The activation function is an integral part of a neural network and does the nonlinear transformation (ie, it describes the input and output relations in a nonlinear way). However, it is this nonlinearity element that allows for higher flexibility and performing complex tasks during the whole model learning process. It helps to speed up the whole learning process. Several activation functions such as sigmoid or ReLU are commonly used in practice.

Sigmoid Function

This function takes a real-value input and converts it into a range between 0 and 1. The sigmoid function is defined as follows:

$$\sigma(x) = 1/(1+e^{-x}) \quad \textbf{(1)}$$

Here it is clear that it will convent the output between 0 and 1 when the input varies in (–∞, ∞). A neuron can use the sigmoid for computing the nonlinear function σ(y = wx + b). If y=wx + b is very large and positive, then $e^{-y} \rightarrow 0$, so $\sigma(y) \rightarrow 1$, whereas

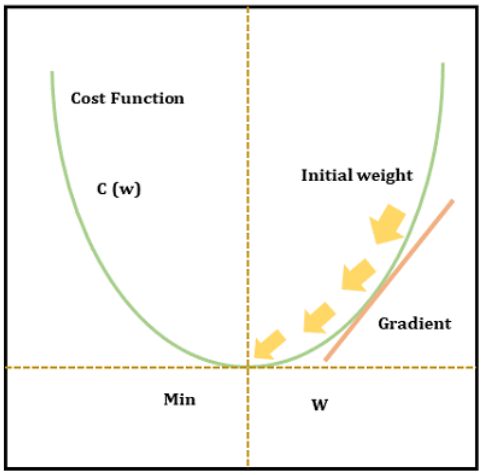
if y = wx + b is very large and negative, then $e^{-y} \rightarrow \infty$, so $\sigma(y) \rightarrow 0$.

ReLU

It is called the rectified linear unit and takes a real input variable and thresholds it at zero (ie, replace native values with zero). The ReLU function is defined as follows:

$$f(x) = \max(0, x) \quad \textbf{(2)}$$

Figure 3. The process of gradient descent.



The cost function C is the initial value and the desired point is C_{min}. The starting weight is w_0, with each step presented as r while the gradient represents the direction of maximum increase. The direction of the value can be expressed mathematically as the partial derivative $ac/\partial w$ to evaluate the time needed for w to reach step r, whereas the opposite direction can be expressed as $-(ac/\partial w)(w_r)$. The most commonly used optimizers are Momentum, Adagrad, AdaDelta, Adam.

Performance Evaluation

We assessed the performance of the DNN model on the validations set for laboratory test recommendations using the following metrics.

Micro-AUC

This averages the prediction matrix. S_{micro} corresponds to a set of correct quadruples. The formula for calculating micro–area under the curve (micro-AUC) is

$$\text{Micro-AUC} = (|S_{micro}|)/[(\Sigma^{m}_{i=1}|Y^{+}_{i.}|)\cdot(\Sigma^{m}_{i=1}|Y^{-}_{i.}|)] \quad \textbf{(3)}$$

$$\text{Smicro} = \{(a,b,i,j)|(a,b)\in Y^{+}_{.i} \times Y^{-}_{.j},\ f_i(x_a) \geq f_j(x_b)\} \quad \textbf{(4)}$$

Macro-AUC

This averages each label. S_{micro} corresponds to a set of correctly ordered instance pairs on each label. The formula for calculating macro-AUC is

$$\text{Macro-AUC} = (1/l)\Sigma^{l}_{j=1}(|S^{j}_{macro}|)/(|Y^{+}_{.j}||Y^{-}_{.j}|) \quad \textbf{(5)}$$

$$S^{j}_{macro} = \{(a,b)\}\in Y^{+}_{.j} \times Y^{-}_{.j}|f_i(x_a) \geq f_i(x_b)| \quad \textbf{(6)}$$

Optimization Algorithms

These algorithms are generally used to minimize errors and generate slightly better and faster results by updating input parameters such as weight and bias values. *Gradient descent* is the most widely used optimization algorithm that helps us understand whether the function is decreasing or increasing at a particular point (Figure 3).

Micro-F1

This averages the prediction matrix, and is calculated as follows:

$$\text{Micro-F1} = (2\Sigma^{l}_{j=1}\Sigma^{m}_{i=1}y_{ij}h_{ij})/(\Sigma^{l}_{j=1}\Sigma^{l}_{i=1}y_{ij} + \Sigma^{l}_{j=1}\Sigma^{m}_{i=1}h_{ij}) \quad \textbf{(7)}$$

Macro-F1

It averages each label, and is calculated as follows:

$$\text{Macro-F1} = (1/l)\Sigma^{l}_{j=1}(2\Sigma^{m}_{i=1}y_{ij}h_{ij})/(\Sigma^{m}_{i=1}y_{ij} + \Sigma^{m}_{i=1}h_{ij}) \quad \textbf{(8)}$$

Average Precision

This reflects the average fraction of relevant labels ranked higher than one other relevant label, and is calculated as follows:

$$\text{Average precision} = (1/m)\Sigma^{m}_{i=1}[1/(|y_{i.}^{+}|)\Sigma_{j\in Y^{+}i.}[|S^{ij}_{precision}|/\text{rankF}(x_i,j)] \quad \textbf{(9)}$$

$$S^{ij}_{precision} = \{k \in Y^{+}_{i.}|\text{rank}_F(x_i,k) \leq \text{rank}_F(x_i,j)|\} \quad \textbf{(10)}$$

Hamming Loss

It is the most commonly used metric to evaluate the performance of a multilabel classifier. It is the average symmetric difference between a set of true labels and a set of predicted labels of the data set. Its formula is as follows:

$$\text{hloss (H)} = (1/ml)\Sigma^{m}_{i=1}\Sigma^{l}_{j-1}[[h_{ij}\neq y_{ij}]] \quad \textbf{(11)}$$

The hamming loss (HL) value ranges from 0 to 1. A lesser value of HL indicates a better classifier.

Results

Prescriptions

In this study, we considered all patients who visited the cardiology department. A total of 37,890 patients visited the department at least once between January 1, 2013, and December 31, 2013. The number of male patients was higher than the number of female patients (51.11% [19,366/37,890] vs 48.89% [18,524/37,890]) and the age of patients ranged from 4 to 102 years. A total of 129,938 prescriptions with laboratory tests were ordered in the cardiology department (Table 1).

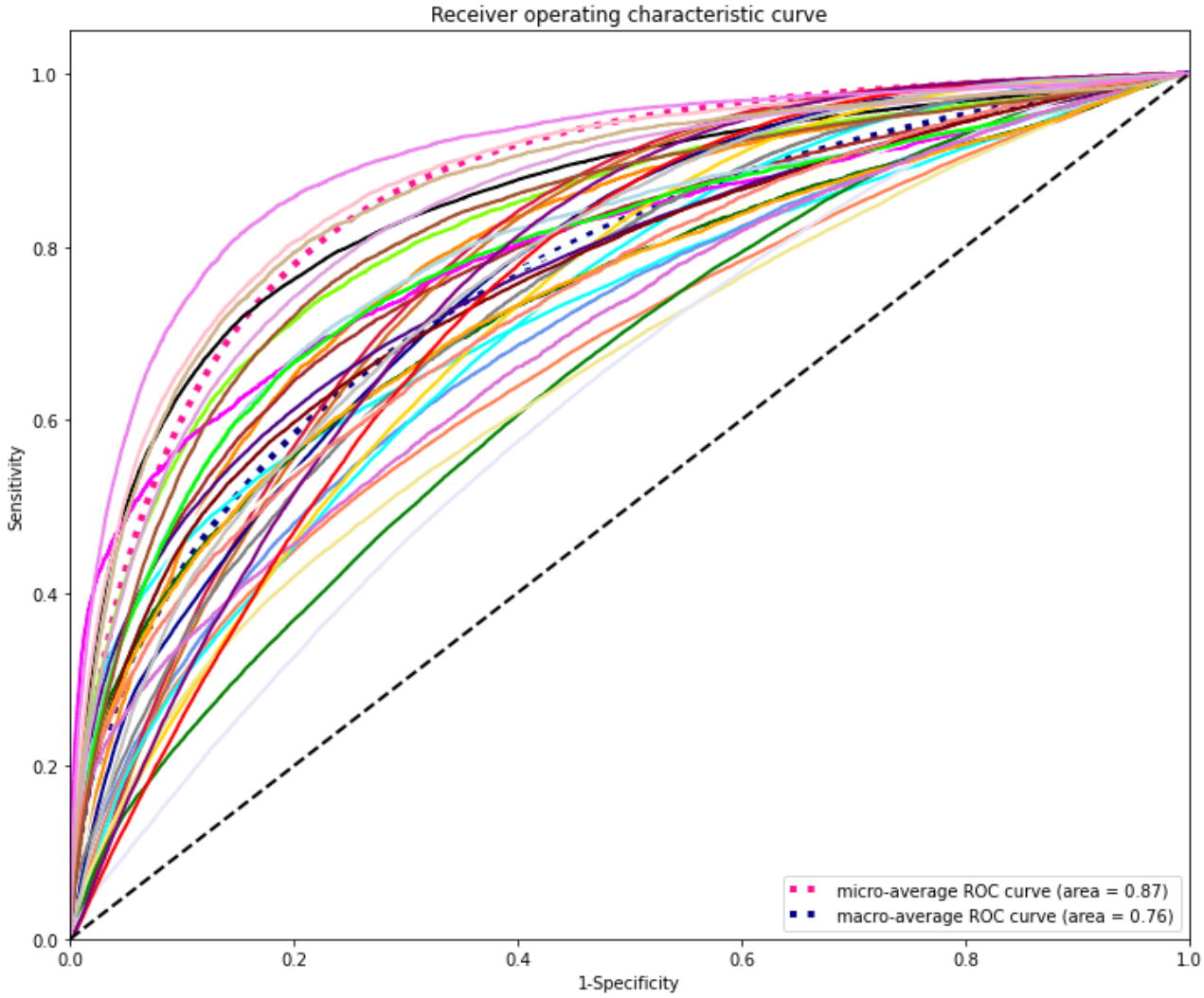
Table 1. Characteristics of patients and clinical variables.

Variables	Values
Total number of prescription	129,938
Total number of patients	37,890
Age (years), range	4-102
Gender	
Male, n (%)	19,366 (51.11)
Female, n (%)	18,524 (48.89)
Number of drugs input	416
Number of diseases input	714
Number of laboratory tests	35

Prediction of Laboratory Tests

A total of 1132 input variables were used to predict the 35 types of laboratory tests. The DL model was applied to data from the cardiology department to predict laboratory tests accurately; the model achieved good discrimination ($AUROC_{macro}$=0.76 and $AUROC_{micro}$=0.87). Figure 4 shows the area under the receiver operating characteristic curve (AUROC) by the DL model. The range of the AUROC was 0.63-0.90 (see Figure S3 and Table S3 in Multimedia Appendix 1).

Figure 4. Receiver operating characteristic (ROC) curves of the deep learning model for predicting laboratory tests.



The DL model's precision, recall, F1 score, and HL based on varying cutoffs for clinical laboratory test prediction are presented in Table 2. Precision, recall, F1 score, and HL ranged from 24% to 56%, 67% to 99%, 36% to 55%, and 0.16 to 0.46, respectively.

Table 2. Recall, precision, F1 score, and hamming loss of the model based on varying cutoffs for clinical laboratory test prediction.

Cutoffs	Recall[a]	Precision[a]	F1 score[a]	Hamming loss
0.01	0.99	0.24	0.36	0.46
0.05	0.94	0.33	0.45	0.39
0.10	0.89	0.40	0.50	0.29
0.15	0.85	0.44	0.52	0.24
0.20	0.80	0.47	0.54	0.21
0.25	0.76	0.51	0.55	0.19
0.30	0.71	0.54	0.55	0.17
0.35	0.67	0.56	0.55	0.16

[a]Overall (micro and macro) result presented.

Discussion

Principal Findings

In this study, we developed and validated a DL-based automated model to recommend laboratory tests based on individual patient clinical history. To our knowledge this is the first study which evaluated the performance of a DL algorithm to recommend laboratory tests, and achieved good performance; therefore, this model can be used in a real-world clinical setting. The main advantage of this model is that it requires minimal input data such as gender, age, disease, and drug information, and thus can be easily integrated into EHR systems. Most importantly, the model can be adjusted for different cutoff values according to physician needs. Moreover, physicians can select the required laboratory tests for an individual patient from a provided list of laboratory tests. This would ensure performing a quick and accurate test. The model showed high discrimination capacity; hence, implementation of this model would ensure accurate laboratory tests, improve patients' safety, and reduce unnecessary costs associated with wrong orders.

Comparison With Other Study

Previously, Wright et al [17] developed an association and data mining technique to suggest laboratory tests. Using a support threshold of 5 and a confidence threshold of 10%, there were 5361 associations between disease and laboratory tests. They reported a higher accuracy (55.6%) across the top 500 associations. The main problems of the system were that the relationships identified were indirect, one to one (drugs and laboratory or disease and laboratory), and had pseudo associations (metformin and hypertension). Furthermore, if the patients had multiple drugs, diseases, and laboratory tests, then their model was unable to find solutions and creates many associations. However, in a real clinical setting, one patient can be suggested to undergo multiple laboratory tests in 1 prescription. For example, a patient with diabetes could have been given multiple tests at the same time. Our model showed a higher accuracy (0.85) to predict laboratory tests. Moreover, our multilabel prediction model could provide a list of laboratory tests based on patients' clinical history; therefore, physicians could choose laboratory tests from provided lists.

Clinical Implications

Health care budgets worldwide are facing increasing pressure to minimize costs while maintaining quality care and ensuring patients' safety [18]. The laboratory tests are often considered a central part of controlling health expenditures and ensuring patients' safety. A previous study in the UK reported that pathology investigations are involved in 70%-80% of all health care decisions and cost the UK National Health Service (NHS) £2.5 billion (US $4 billion) annually [19]. Proper utilization of limited resources and curbing numerous unnecessary laboratory tests will help reduce health care costs because approximately 2.9%-56% of all laboratory tests are reported to be likely overutilized [20]. About 30% of the outpatient laboratory tests were found to be inappropriate and ordered just for patient check-ups [21]. However, inappropriate testing can contribute to increasing patient anxiety, iatrogenic anemia, and patient dissatisfaction.

Several groups of researchers have proposed many ways to control inappropriate laboratory test ordering, but it remains unclear which is the most effective or how to integrate these ways with other systems designed to control laboratory costs. Some have suggested reducing the reimbursement rate to control expenditures on laboratory services. Although this approach can be effective in the short run, it has several fundamental flaws. The second approach is linked to medical necessity; laboratory test cost can be reduced by decreasing the utilization of tests that are not medically necessary; however, it is very difficult to define the appropriate use of laboratory tests. Albeit significant progress has already been made, much work remains to be done in this area. A third approach has been active management of test utilization by laboratory staff. This approach has been used mostly in academic medical centers, often integrated as part of training for residents and fellows [22]. It can also vary from simply having laboratory staff act as gatekeepers for specific tests to more robust, systematic methods for improving test utilization [4]. However, our automated laboratory tests recommendation system is based on real-world clinical data and showed high discrimination capabilities. This model can thus help reduce unnecessary health care costs by recommending exact and appropriate laboratory tests based on individual patient problems.

Strengths and Limitations

Our study has several strengths that need to be addressed. (1) This is the first study to evaluate and utilize the DL model to recommend laboratory tests using variables available in EHR. This study can therefore be used as a benchmark for future studies. (2) Our novel model is significantly more accurate and can adjust the cutoff value according to physician demand. Third, our evaluation of DL algorithms was rigorous, including fewer variables, and the model was developed based on daily clinical practice data.

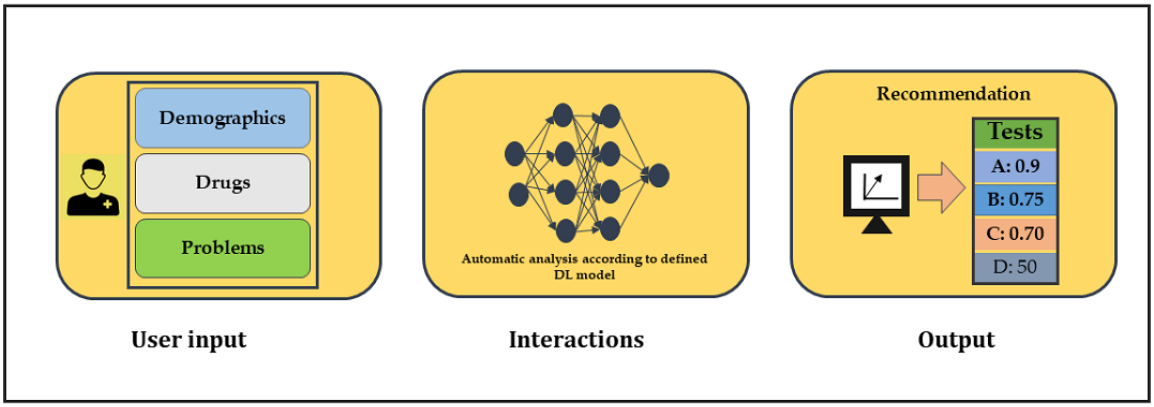
Our study has also several limitations: (1) Our model was developed based on data from only the cardiology department; however, this model can be extended for use in other departments using their own data. (2) This model used only 35 laboratory tests in the prediction model; however, it covered more than 90% of total tests ordered. (3) Our model has not yet been tested using an external data set; however, we used internal validation to evaluate model performance. Sometimes performance of the model may deviate when it is validated using other data sets but this is not to a large extent.

Future Perspective

Our next step is to extend this work to other departments and includes using nonstructured data such as progress notes and operative notes. We believe that the inclusion of these data could increase our model performance. Moreover, we will use 10-year data to improve our model performance, although it would be computationally expensive. We also have a plan to include procedures in our system because it would further add value in the real-world clinical setting. Because our model showed higher sensitivity and a less false-positive rate, we will integrate our model with EHR to improve clinical decisions and reduce laboratory error rates. Although this would be quite powerful, it remains challenging for several reasons, including

gold-standard evaluation and the acceptability of our model in clinical settings. However, one potential benefit of implementing this model in real-world clinical settings would be individual physician selection choice from a list of provided laboratory tests based on probability (Figure 5). It would not trigger many alerts and not hamper workflow.

Figure 5. Proposed infographic of deep learning (DL)-based laboratory testing recommendation tool.



Conclusion

Using commonly available clinical variables, we developed and validated a DL algorithm that predicts laboratory tests with high accuracy, and recommends clinically relevant laboratory tests at the time of ordering. To our knowledge, this is the first study to evaluate the performance of algorithms and this predictive algorithm can serve as a clinical decision-support tool. Most importantly, our model could help reduce unnecessary laboratory test ordering and health care costs. The integration of this model into daily clinical practice may facilitate optimal laboratory test selection based on the appropriate thresholds. However, further research is necessary to assess the workflow of the system, and weigh the benefits of patients and physicians while implementing the model as an effective recommendation tool in clinical practice.

Acknowledgments

This research is funded in part by the Ministry of Education (MOE) under grant MOE 109-6604-001-400 and DP2-109-21121-01-A-01 and the Ministry of Science and Technology (MOST) under grant MOST 109-2823-8-038-004. We thank our colleagues who edited our manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Artificial intelligence-based automated recommendation system for clinical laboratory tests.
[DOCX File , 441 KB - medinform_v8i11e24163_app1.docx]

References

1. Dahm MR, Georgiou A, Westbrook JI, Greenfield D, Horvath AR, Wakefield D, et al. Delivering safe and effective test-result communication, management and follow-up: a mixed-methods study protocol. BMJ Open 2018 Feb 15;8(2):e020235. [doi: 10.1136/bmjopen-2017-020235] [Medline: 29449297]
2. Zhi M, Ding EL, Theisen-Toupal J, Whelan J, Arnaout R. The landscape of inappropriate laboratory testing: a 15-year meta-analysis. PLoS One 2013 Nov 15;8(11):e78962 [FREE Full text] [doi: 10.1371/journal.pone.0078962] [Medline: 24260139]
3. Hammerling J. A Review of Medical Errors in Laboratory Diagnostics and Where We Are Today. Lab Med 2012 Feb 01;43(2):41-44 [FREE Full text] [doi: 10.1309/LM6ER9WJR1IHQAUY]
4. Kim JY, Dzik WH, Dighe AS, Lewandrowski KB. Utilization Management in a Large Urban Academic Medical Center. Am J Clin Pathol 2011 Jan 01;135(1):108-118. [doi: 10.1309/ajcp4gs7ksbdbacf]
5. Robinson A. Rationale for cost-effective laboratory medicine. Clin Microbiol Rev 1994 Apr 01;7(2):185-199. [doi: 10.1128/cmr.7.2.185] [Medline: 8055467]
6. Stuebing EA, Miner TJ. Surgical vampires and rising health care expenditure: reducing the cost of daily phlebotomy. Arch Surg 2011 May 01;146(5):524-527. [doi: 10.1001/archsurg.2011.103] [Medline: 21576605]
7. May TA, Clancy M, Critchfield J, Ebeling F, Enriquez A, Gallagher C, et al. Reducing Unnecessary Inpatient Laboratory Testing in a Teaching Hospital. Am J Clin Pathol 2006 Aug 01;126(2):200-206. [doi: 10.1309/wp59ym73l6cegx2f]

8. Driskell O, Holland D, Hanna F, Jones P, Pemberton R, Tran M, et al. Inappropriate requesting of glycated hemoglobin (Hb A1c) is widespread: assessment of prevalence, impact of national guidance, and practice-to-practice variability. Clin Chem 2012 May;58(5):906-915. [doi: 10.1373/clinchem.2011.176487] [Medline: 22344287]
9. Wu C, Hsu W, Islam MM, Poly TN, Yang H, Nguyen P, et al. An artificial intelligence approach to early predict non-ST-elevation myocardial infarction patients with chest pain. Comput Methods Programs Biomed 2019 May;173:109-117. [doi: 10.1016/j.cmpb.2019.01.013] [Medline: 31046985]
10. Islam M, Yang H, Nguyen P, Wang Y, Poly T, Li YCJ. Deep Learning Approach for the Development of a Novel Predictive Model for Prostate Cancer. Stud Health Technol Inform 2020 Jun 16;270:1241-1242. [doi: 10.3233/SHTI200382] [Medline: 32570599]
11. Islam MM, Yang H, Poly TN, Jian W, Jack Li YC. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. Comput Methods Programs Biomed 2020 Jul;191:105320. [doi: 10.1016/j.cmpb.2020.105320] [Medline: 32088490]
12. Gupta R, Pal S, Kanade A, Shevade S. DeepFix: Fixing common C language errors by deep learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Menlo Park, CA: Association for the Advancement of Artificial Intelligence; 2017 Presented at: The Thirty-First AAAI Conference on Artificial Intelligence; February 4–9; San Francisco, CA p. 1345-1351 URL: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14603/13921
13. Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: An overview. New York: IEEE; 2013 May Presented at: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; May 26-31, 2013; Vancouver, BC p. 8599-8603 URL: https://ieeexplore.ieee.org/document/6639344 [doi: 10.1109/icassp.2013.6639344]
14. Hsing AW, Ioannidis JPA. Nationwide Population Science: Lessons From the Taiwan National Health Insurance Research Database. JAMA Intern Med 2015 Sep 01;175(9):1527-1529. [doi: 10.1001/jamainternmed.2015.3540] [Medline: 26192815]
15. Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing. JAMA 2020 Apr 14;323(14):1341-1342. [doi: 10.1001/jama.2020.3151] [Medline: 32125371]
16. Tsai C, Yang H, Islam M, Hsieh WS, Juan SH, Chen JC, et al. Psychotropic medications prescribing trends in adolescents: A nationwide population-based study in Taiwan. Int J Qual Health Care 2017 Oct 01;29(6):861-866. [doi: 10.1093/intqhc/mzx123] [Medline: 29036295]
17. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 2010 Dec;43(6):891-901 [FREE Full text] [doi: 10.1016/j.jbi.2010.09.009] [Medline: 20884377]
18. Fryer AA, Smellie WSA. Managing demand for laboratory tests: a laboratory toolkit. J Clin Pathol 2013 Jan 26;66(1):62-72. [doi: 10.1136/jclinpath-2011-200524] [Medline: 23015659]
19. Kwok J, Jones B. Unnecessary repeat requesting of tests: an audit in a government hospital immunology laboratory. J Clin Pathol 2005 May 01;58(5):457-462. [doi: 10.1136/jcp.2004.021691] [Medline: 15858114]
20. Kiechle FL, Arcenas RC, Rogers LC. Establishing benchmarks and metrics for disruptive technologies, inappropriate and obsolete tests in the clinical laboratory. Clin Chim Acta 2014 Jan 01;427:131-136 [FREE Full text] [doi: 10.1016/j.cca.2013.05.024] [Medline: 23732401]
21. Baricchi R, Zini M, Nibali MG, Vezzosi W, Insegnante V, Manfuso C, et al. Using pathology-specific laboratory profiles in clinical pathology to reduce inappropriate test requesting: two completed audit cycles. BMC Health Serv Res 2012 Jul 03;12(1):187 [FREE Full text] [doi: 10.1186/1472-6963-12-187] [Medline: 22759353]
22. Aesif S, Parenti D, Lesky L, Keiser J. A cost-effective interdisciplinary approach to microbiologic send-out test use. Arch Pathol Lab Med 2015 Feb;139(2):194-198 [FREE Full text] [doi: 10.5858/arpa.2013-0693-OA] [Medline: 24758733]

Abbreviations

ACE: angiotensin-converting enzyme
AUC: area under the curve
AUROC: area under the receiver operating characteristic curve
DL: deep learning
DNN: deep neural network
EHR: electronic health record
HL: hamming loss
NHIRD: National Health Insurance Research and Development
NHS: National Health Service

Edited by G Eysenbach; submitted 07.09.20; peer-reviewed by Anonymous, M Tourey; comments to author 21.09.20; revised version received 28.09.20; accepted 30.09.20; published 18.11.20.

Please cite as:
Islam MM, Yang HC, Poly TN, Li YCJ
Development of an Artificial Intelligence–Based Automated Recommendation System for Clinical Laboratory Tests: Retrospective Analysis of the National Health Insurance Database
JMIR Med Inform 2020;8(11):e24163
URL: https://medinform.jmir.org/2020/11/e24163
doi:10.2196/24163
PMID:33206057

Original Paper

Machine Learning Approach to Reduce Alert Fatigue Using a Disease Medication–Related Clinical Decision Support System: Model Development and Validation

Tahmina Nasrin Poly[1,2,3], MSc; Md.Mohaimenul Islam[1,2,3], MSc; Muhammad Solihuddin Muhtar[2], BSc; Hsuan-Chia Yang[1,2,3], PhD; Phung Anh (Alex) Nguyen[2,4], PhD; Yu-Chuan (Jack) Li[1,2,3,5,6], MD, PhD

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

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

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

[4]Department of Healthcare Information & Management, Ming Chuan University, Taoyuan City, Taiwan

[5]Department of Dermatology, Wan Fang Hospital, Taipei, Taiwan

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

Corresponding Author:
Yu-Chuan (Jack) Li, MD, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
15 F, No. 172-1, Sec. 2, Kellung Rd, Da'an dist
Taipei, 106
Taiwan
Phone: 886 0966546813
Email: jaak88@gmail.com

Abstract

Background: Computerized physician order entry (CPOE) systems are incorporated into clinical decision support systems (CDSSs) to reduce medication errors and improve patient safety. Automatic alerts generated from CDSSs can directly assist physicians in making useful clinical decisions and can help shape prescribing behavior. Multiple studies reported that approximately 90%-96% of alerts are overridden by physicians, which raises questions about the effectiveness of CDSSs. There is intense interest in developing sophisticated methods to combat alert fatigue, but there is no consensus on the optimal approaches so far.

Objective: Our objective was to develop machine learning prediction models to predict physicians' responses in order to reduce alert fatigue from disease medication–related CDSSs.

Methods: We collected data from a disease medication–related CDSS from a university teaching hospital in Taiwan. We considered prescriptions that triggered alerts in the CDSS between August 2018 and May 2019. Machine learning models, such as artificial neural network (ANN), random forest (RF), naïve Bayes (NB), gradient boosting (GB), and support vector machine (SVM), were used to develop prediction models. The data were randomly split into training (80%) and testing (20%) datasets.

Results: A total of 6453 prescriptions were used in our model. The ANN machine learning prediction model demonstrated excellent discrimination (area under the receiver operating characteristic curve [AUROC] 0.94; accuracy 0.85), whereas the RF, NB, GB, and SVM models had AUROCs of 0.93, 0.91, 0.91, and 0.80, respectively. The sensitivity and specificity of the ANN model were 0.87 and 0.83, respectively.

Conclusions: In this study, ANN showed substantially better performance in predicting individual physician responses to an alert from a disease medication–related CDSS, as compared to the other models. To our knowledge, this is the first study to use machine learning models to predict physician responses to alerts; furthermore, it can help to develop sophisticated CDSSs in real-world clinical settings.

(JMIR Med Inform 2020;8(11):e19489) doi:10.2196/19489

KEYWORDS

clinical decision support system; alert fatigue; machine learning; artificial neural network

Introduction

Initiation of computerized provider order entry (CPOE) systems has allowed physicians to order medications, laboratory tests, and other ancillary services electronically [1]. CPOE systems create an opportunity to improve patient care by decreasing medication errors, reducing redundant test orders, and promoting standardized clinical practice [2,3]. However, CPOE is often integrated with a clinical decision support system (CDSS) in order to make better clinical decisions through guidance, alerts, and reminders. A CDSS is always combined with software algorithms that generate alerts during orders entered into a CPOE by physicians [4,5]. Each of these alerts addresses a meaningful clinical issue relevant to the administration process and has a positive impact on identifying and preventing erroneous or less optimal prescription [6-8].

The productivity of CDSSs is often impaired by generating distracting alerts in the system (ie, a high volume of clinically irrelevant alerts) [9,10]. van der Sijs et al [11] suggested that an ideal CDSS should have high specificity and sensitivity, provide clear information, and facilitate safe and efficient handling of alerts. A recent study reported that approximately 90%-95% of medication alerts are overridden by providers [12,13], and more than half of overrides are due to alerts being deemed clinically irrelevant [14]. The main concern is that these large numbers of clinically irrelevant alerts might cause alert fatigue and consume too much time and mental energy. Moreover, it sometimes leads staff to override both critical warnings and unimportant alerts. Getting frequent false alerts can desensitize physicians so that providers always ignore and mistrust alerts with acceleration [15]. Ignoring clinically relevant alerts too much triggers patient harm and is associated with an increased rate of mortality.

Until now, significant efforts and strategies have been implemented in minimizing alert fatigue, such as the administration of highly specific algorithms [16], customization of third-party providers' sets of alerts [17], and execution of tiered severity grading to stratify and lessen the number of false alerts [18]. Several studies suggested turning off frequently overridden alerts [19], updating clinical content to deliver the most current evidence at the point of care, and holding consensus meetings between physicians and pharmacists [20]. Since physicians increasingly adopt electronic prescribing, the progression and proclamation of CDSS alerts might depend, in part, on whether providers find medication safety alerts valuable.

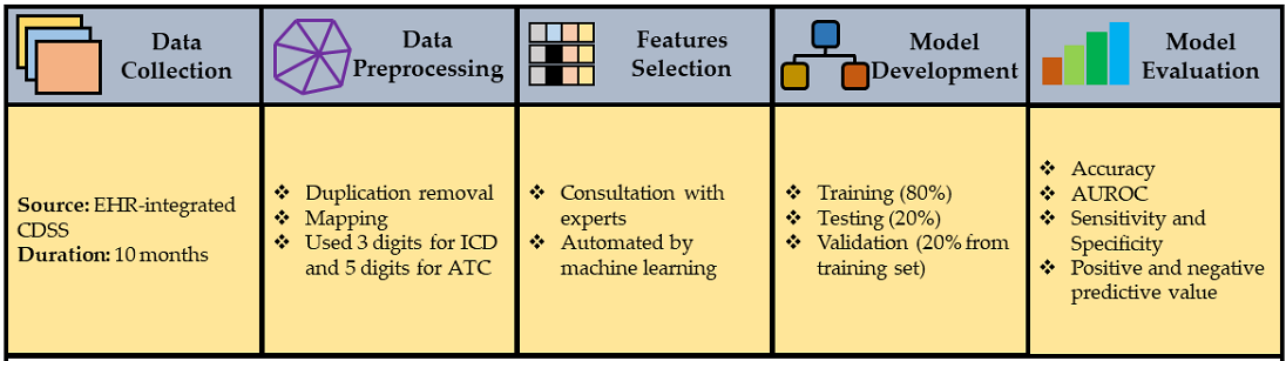
Machine learning is comprised of a collection of techniques that has the potential to learn complex rules and to identify patterns from multidimensional datasets. It has been effectively employed in many areas, such as disease risk prediction [21], classification [22], and health care utilization [23]. To our knowledge, no studies have examined machine learning techniques regarding medication alert reduction in a large number of alert analyses among physicians of different specialties. We hypothesized that machine learning models could predict physician responses, which would ultimately directly assist in developing a sophisticated CDSS for reducing alert fatigue. Therefore, the primary objective of this study was to develop and validate machine learning models to reduce alert fatigue by predicting physician responses. This study may provide perspective on the perceived usefulness of CDSS alerts in patient care and insights into how to design better alert systems in real-world clinical settings. It can contribute to minimizing the number of alerts in the user interface, ensuring the appropriate prescription, and reducing the severity of unintended consequences.

Methods

Ethical Approval and Study Process

This type of study does not require Institutional Review Board review, following the policy of the National Health Research Institutes in Taiwan, as it provides a large amount of computerized, deidentified data. The entire study process is shown in Figure 1.

Figure 1. Study design process. ATC: Anatomical Therapeutic Chemical classification system; AUROC: area under the receiver operating characteristic curve; CDSS: clinical decision support system; EHR: electronic health record; ICD: International Classification of Diseases.



Data Source

We collected data from an electronic health record (EHR)–integrated disease medication–related CDSS from a university teaching hospital in Taiwan. We considered only prescriptions that generated alerts due to a prescription error in the CDSS. The data collection period was between August 2018 and May 2019. During the 10-month study period, 9213 prescriptions generated alerts that accounted for approximately 3% of total prescriptions provided by physicians.

Data Preprocessing

The first step of this study was to clean the data. In the dataset, lots of duplications of prescriptions appeared, which means there were several prescriptions with the same patient's registration number, diagnosis code (ie, International Classification of Diseases, 10th Revision, Clinical Modification [ICD-10-CM]), and drug code (ie, Anatomical Therapeutic Chemical [ATC] classification system code). Therefore, we removed those prescriptions and kept the most recent prescription. A total of 6453 prescriptions were considered to develop machine learning–based prediction models. A prescription with the Taiwan National Health Insurance code as the diagnosis code was mapped to the ICD-10-CM code. Data normalization was carried out by converting all the values between 0 and 1. Finally, the data were converted into a matrix that included the diagnosis code, drug code, department ID, and physician ID.

Feature Selection

There could be more than 20 different clinical variables available in a single prescription. Therefore, feature selection is essential in order to keep the variables within a manageable size to be able to optimize the prediction model. The feature selection process was completed in three stages: (1) consultation with an expert (YL) who is a physician and specialist in CDSSs, (2) automated feature selection via machine learning algorithms, and (3) reduction of the number of input variables by using only the first three digits of the diagnosis code (ie, ICD-10-CM) and the first five digits of the drug code (ie, ATC). The patient's age, the patient's gender, the diagnosis code (ie, ICD-10-CM), the drug code (ie, ATC), the physician ID, and the department ID were considered as input variables. We then created a matrix for the diagnosis code (ie, ICD-10-CM), the drug code (ie, ATC code), the physician ID, and the department ID. A total of 6453 input variables were used to develop a machine learning model with binary outcomes.

Table 1. List of input variables.

Variable	Input column contents	Input column number
Patient's gender	Male or female	1
Diagnosis code	First 3 digits of the ICD-10-CM[a] code	822
Drug code	First 5 digits of the ATC[b] classification system code	262
Physician ID	Physician ID	227
Department ID	Department ID	29

[a]ICD-10-CM International Classification of Diseases, 10th Revision, Clinical Modification.

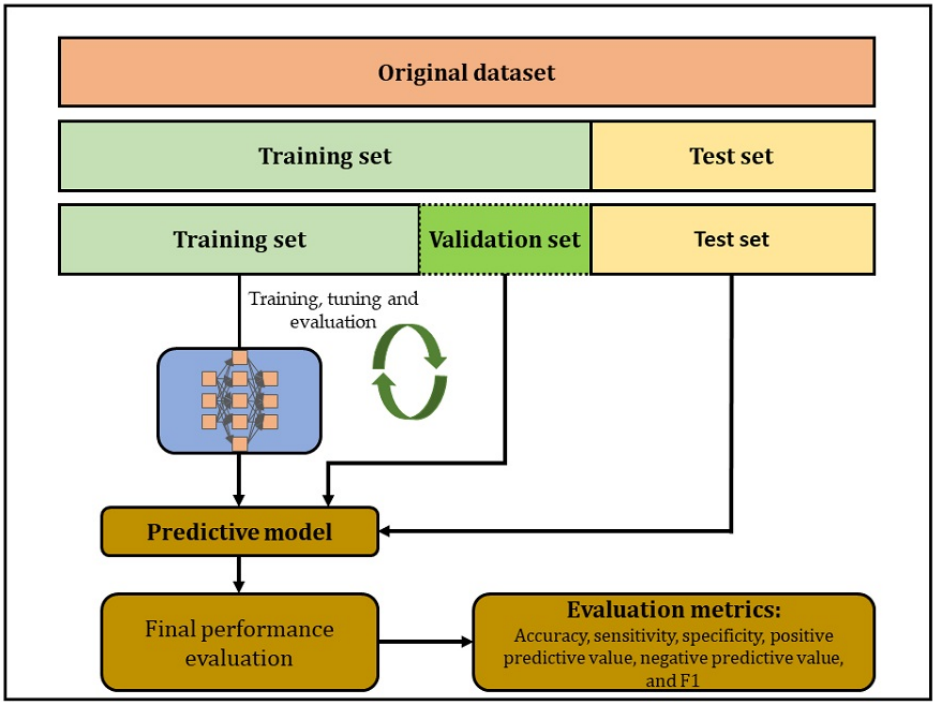
[b]ATC: Anatomical Therapeutic Chemical.

Model Development

Overview

The objective of the model was to reliably predict what would be physicians' responses to an alert. We divided the entire dataset into two parts: (1) the training dataset (80% of the dataset) and (2) the testing dataset (20% of the dataset). However, the model was trained using 60% of the dataset as the training set and 20% of the dataset as an internal validation set. The remaining 20% of the dataset was used for testing our model's performance (see Figure 2). Model development was carried out using Python 3.6 software (Python Software Foundation). Python is a free and open-source programming language and environment for statistical computing and graphics.

Figure 2. Distribution of training and testing datasets for model development.



Artificial Neural Network

Artificial neural networks (ANNs) were first introduced in the 1940s; recently, they have become extremely powerful and one of the most popular machine learning models that interconnects with adaptive simple processing elements. They usually work by mimicking the biological nervous systems responsible for knowledge processing and knowledge representation [24]. ANN-based algorithms have already shown high performance in terms of accuracy, sensitivity, and specificity for classification problems. Therefore, the application of ANNs has increased globally in recent years in health care research, including in drug development, pattern recognition, disease prediction, disease diagnosis, and disease prognosis. ANNs consist of three layers of neurons: the *input layer*, the *hidden layers*, and the *output layer*. The hidden layer can be a single or multiple layer. Every hidden layer is comprised of an activation function. In our study, we used three hidden layers, with the *rectified linear unit* (ReLU) activation function in the first and second hidden layers, and the *sigmoid* activation function in the third hidden layer.

The ReLU is a widely used activation function in the prediction model. It converts input values from 0 to α. In the third layer, we used a sigmoid activation function due to a nonlinear nature. The sigmoid function is also one of the most commonly used activation functions for binary classification. The sigmoid function converts output classes between 0 and 1. The ANN was designed to be a classification model that can predict the responses from multiple physicians while minimizing prediction errors by using *binary cross-entropy* as loss function and the stochastic gradient descent method for optimization. Moreover, 100 epochs were used in the ANN model where maximum accuracy and minimum loss for training and validation can be achieved.

Random Forest

Random forest (RF) is also known as *ensemble learning* because it is an ensemble of a large number of individual decision trees [25]. Each tree in the RF model spits out a class prediction, and the class with the most votes becomes our model's prediction. However, RF applies to both the *classification* and *regression* models.

Naïve Bayes

Naïve Bayes (NB) is a classification model that uses the *Bayesian probability* theory during prediction [26]. It is also known as a probabilistic classifier. In 1960, the NB model was first introduced for text classification by the text retriever community [27]. However, there are several types of NB algorithms for parameter estimation and event models, such as *Gaussian naïve Bayes*, *multinomial naïve Bayes*, and *Bernoulli naïve Bayes*. Bayes theorem is expressed as equation 1 in Figure 3.

Figure 3. Equations. FN: false negative; FP: false positive; NPV: negative predictive value; PPV: positive predictive value; TN: true negative; TP: true positive.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (1)$$

$$X = (x_1 + x_2 + x_3, \ldots \ldots, x_n) \quad (2)$$

$$P(y|x_i, \ldots., x_n) = \frac{P(x_i|y)P(x_2|y)\ldots P(x_n|y)\,P(y)}{P(x_1)\,P(x_2)\ldots P(x_n)} \quad (3)$$

$$y = argmax_y\, P(y) \prod_{i=1}^{n} P(x_i|y) \quad (4)$$

$$\hat{A} = \arg\min E_{x,y}\left[L(y, A(x))\right] \quad (5)$$

$$\hat{A}(x) = \sum_{i=1}^{M} y_1\, h_i(x) + canstant \quad (6)$$

$$h(x_i) = \begin{cases} +1 & if\ w.x + b \geq 0 \\ -1 & if\ w.x + b < 0 \end{cases} \quad (7)$$

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i(w.x_i - b)\right)\right] + \lambda\|w\|^2 \quad (8)$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (9)$$

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (10)$$

$$Specificity = \frac{TN}{(TN+FP)} \quad (11)$$

$$PPV = \frac{(Sensitivity \times \pi)}{(Sensitivity \times \pi + (1 - Specificity) \times (1-\pi))} \quad (12)$$

$$NPV = \frac{(Specificity \times (1-\pi))}{(Specificity \times (1-\pi) + (1 - Sensiv\,ity) \times \pi)} \quad (13)$$

The variable *y* is the class variable that represents whether the alert will be accepted or rejected given the condition. Variable *X* represents the features like drugs, disease, and demographic. *X* is given as equation 2 in Figure 3.

Here, x_1, x_2 ... x_n represent the features (ie, they can be mapped to outcome: accept or reject alert). By substituting for *X* and expanding using the chain, the rule is given in equation 3 in Figure 3. In our model, the class variable *y* has two outcomes: accept or reject. There could be cases where the classification is multivariable. Therefore, the equation 4 in Figure 3 is used to find the class variable *y* with maximum probability.

Gradient Boosting

Gradient boosting (GB) is one of the promising machine learning algorithms that has already shown better prediction for classification [28]. It can be used both in classification and regression models. Like RF, GB is a set of decision trees, but the main differences are how the trees are built and how the results are combined. In the RF model, each tree is built independently, while in the GB model they are built one tree at a time. The GB model works in a forward stage-wise manner and converts weak learners to strong learners [29]. The most interesting part of the GB algorithm is that it can easily fit into the new model. Moreover, the RF model combines results at the end of the process, by averaging or *majority rules*, while the GB model combines results along the way [30].

In the training set, input variables such as drugs and diseases, make a set $\{(x_1y_1), \ldots, (x_ny_n)\}$ of known values of *x* and corresponding values of *y*. The goal is always to find an approximation $\hat{A}(x)$ to find a function $A(x)$ that minimizes the expected value of the specified loss function $L(y, A(x))$, as shown in equation 5 in Figure 3.

The GB model assumes a real-valued *y* and calculates an approximation $\hat{A}(x)$ in the form of a weighted sum of functions $h_i(x)$ from *H* classes, which are called base or weak learners, as shown in equation 6 in Figure 3.

Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm. SVM is used both in *classification* and *regression* problems [31]. It is also used to solve linear and nonlinear problems and works well for many complex problems. The idea of SVM is simple: it creates a line or a hyperplane that separates the data into classes. The hypothesis function *h* is defined as shown in equation 7 in Figure 3.

The point above or on the hyperplane is classified as a class +1, and the point below the hyperplane is classified as a class –1. The SVM classifier works in the form shown in equation 8 in Figure 3.

Model Performances

Overview

To evaluate the performance of five machine learning algorithms, we calculated accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the area under the receiver operating characteristic curve (AUROC). For calculating those measures, we measured true positive, true negative, false positive, and false negative. The definitions of the six parameters are given below.

Accuracy

Accuracy is the test by which we can see how accurate our model is. The equation to calculate accuracy is shown in equation 9 in Figure 3.

Sensitivity

Sensitivity is the test by which we can determine a positively identified case. The equation to calculate sensitivity is shown in equation 10 in Figure 3.

Specificity

Specificity is the measure by which we can measure correctly identified cases from negative cases. The mathematical equation is given in equation 10 in Figure 3.

PPV and NPV

PPV and NPV are two basic measures in biomedical studies. PPV is the probability that the positively identified case is positive. The mathematical equation for PPV is given in equation 12 in Figure 3. Similarly, the NPV is the probability that the negatively identified case is negative. The mathematical equation is given in equation 13 in Figure 3.

AUROC

AUROC is a performance measure by which we can evaluate the performance of the model. AUROC is a performance matrix for *discrimination*; it shows the predictive model's ability to discriminate between positive and negative cases.

Results

Dataset Characteristics

A total of 9214 prescriptions with an alert were collected during the 10-month study period. After preprocessing and removing duplicate prescriptions with the same registration numbers, 6453 prescriptions were used to develop our models. The neurology department got the highest number of alerts (1039/6453, 16.10%). Of those alerts, 546 (52.55%) were accepted and 493 (47.45%) were rejected by physicians (see Multimedia Appendix 1, Figure S1). The urology, dermatology, chest medicine, family medicine, metabolism, and otolaryngology departments observed higher alert rates of 10.61% (685/6453), 9.80% (633/6453), 6.91% (446/6453), 6.61% (427/6453), 6.52% (421/6453), and 6.50% (420/6453), respectively. Moreover, eight departments, including rehabilitation medicine, infectious disease, and ophthalmology, had alert rates of more than 1%. Gender, diagnosis codes, disease codes, physician IDs, and department IDs were used to develop and validate our prediction model (see Table 1).

Prediction Performance of Machine Learning Algorithms

We developed five types of machine learning models to predict physician response. To determine the overall performance of predictive models, six evaluation metrics were applied. Among all the machine learning models, ANN showed the best performance (AUROC 0.94) (see Figure 4 and Multimedia Appendix 1, Figure S2).

The accuracy of the ANN, RF, NB, GB, and SVM models were 0.88, 0.85, 0.83, 0.82, and 0.57. The sensitivity and specificity of the ANN, RF, NB, GB, and SVM models were 0.87, 0.88, 0.87, 0.79, and 0.57 and 0.83, 0.82, 0.78, 0.90, and 1.0, respectively (see Table 2).

Figure 4. Performance of machine learning prediction models. ANN: artificial neural network; GB: gradient boosting; RF: random forest; NB: naïve Bayes; ROC: receiver operating characteristic; SVM: support vector machine.

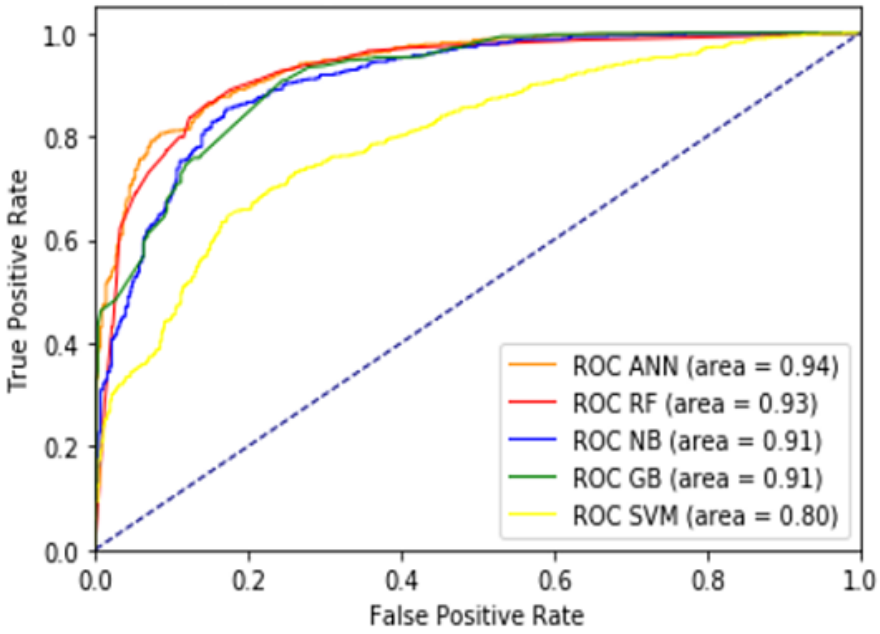


Table 2. Performance of the prediction models.

Algorithm	Accuracy	Sensitivity, %	Specificity, %	PPV[a], %	NPV[b], %	F1
Artificial neural network	0.885	87.01	83.46	87.84	82.40	87.42
Random forest	0.857	88.29	82.48	86.62	84.57	87.44
Naïve Bayes	0.835	87.48	78.74	83.22	84.03	85.29
Gradient boosting	0.828	79.46	90.24	94.59	67.15	86.36
Support vector machine	0.575	57.45	100.0	100.0	0.54	72.97

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

Discussion

Principal Findings

CDSSs directly assist physicians in making correct clinical decisions that ultimately reduce prescription errors by generating real-time alerts and lessen probable unwanted consequences. Clinical workflow is often impaired by excessive numbers of alerts; therefore, physicians pay less attention to alerts and even ignore alerts indiscriminately. This study focused on physicians' recent practice patterns and represented the findings of machine learning models to predict physicians' responses to alerts from a disease medication–related CDSS. The key findings are as follows: (1) an ANN model can correctly predict physicians' responses with higher accuracy than other models and (2) we identified potential features that could provide insight into the system design. These findings may contribute to building a sophisticated provider-friendly interface in which a CDSS may offer real-time alerts if the prediction is positive for that individual physician. If the prediction is negative, that means physicians might not accept the alert; therefore, the CDSS will not generate alerts during the prescribing of prescriptions or will provide soft or passive alerts without interruption. However, all the alerts would be recorded and the report sent to the individual physician by email on a weekly basis to inform them of how important the alerts were in order to reduce unwanted consequences.

Clinical Implications

CDSSs have already shown their capability to improve patient safety and quality of care by lowering the number preventable medication errors [32-34]; however, an unreasonable override rate raises questions regarding the quality of CDSSs. Patient safety and effective care could be improved by initiating sophisticated criteria for generating alerts in the CDSS that prevent alert fatigue and minimize the override rate [35-37]. Identification of physicians and departments who override alerts more often would help to reduce the override rate and help us understand how physicians would respond to drug-disease alerts, which would result in immense benefits. There are no previous studies that used a machine learning prediction model to identify physicians and departments who override alerts more often. In

this study, machine learning algorithms were used to reduce alert fatigue by identifying physicians and departments who override alerts more often. Our findings are consistent with existing research that showed physicians played a great role in alert override [38]. Bell et al showed that alert override can be minimized by physicians' preferences for alert selection [39]. There are several reasons that can make physicians override alerts. First, current medication-related CDSSs are not designed to take the patient's previous medication history into account. Sometimes patients are already tolerant of the drug and physicians need to override the alert and prescribe the drug [40]. Second, some CDSSs required an entry for the reason for alert override and that lead to an unacceptable time burden for physicians [41]. Third, physicians believe that *they already know* the alert is inappropriate based on their experience, so they are more likely to override the alert [13,42]. Our study also provides a very important point: no matter how accurate the CDSS is according to the most relevant knowledge base, the alert acceptance was highly affected by the individual physician's perspective. Our model will reduce the gap between real-world clinical practice and knowledge-based theory.

Yeh et al [43] demonstrated that dermatology, gynecology-obstetrics, family medicine, and ophthalmology departments had higher acceptance rates; however, pediatrics, psychiatry, and internal medicine departments, such as cardiology, endocrinology and metabolism, gastroenterology, hematology, rheumatology, and general medicine, had lower acceptance rates. In our study, we also found that physicians' decisions vary from department to department.

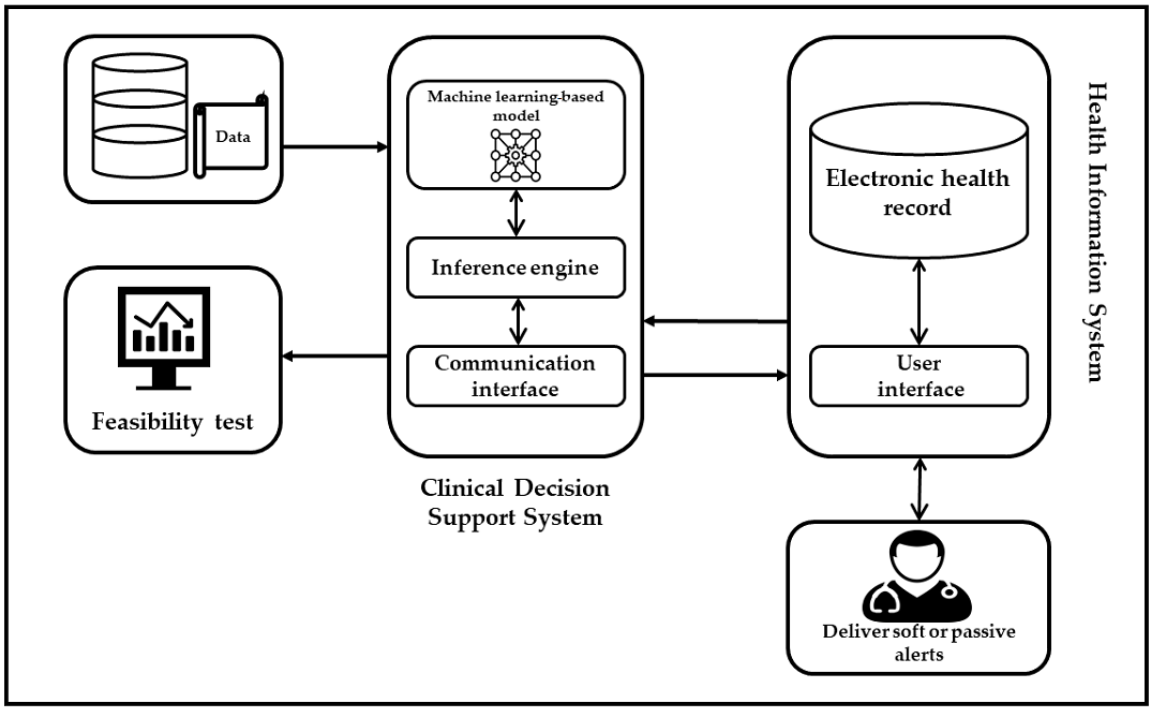
Strengths and Limitations

This study has several strengths. First, this is the first study to use machine learning algorithms to predict physicians' intentions to accept or reject alerts. This model may help to reduce alert fatigue in the current CDSS. Second, this study is personalized for each physician. Third, the performance of the model is satisfactory, such that it would help to reduce alert fatigue. Despite several strengths, our study also has several limitations that need to be addressed. First, we did not include free-text override reasons in our analysis, and free-text reasons could add additional value to our model. However, our model provided the AUROC with decent specificity and sensitivity. Second, we did not include physicians' experiences, working periods, ages, and genders in this prediction model. These data are difficult to collect retrospectively because EHR systems do not record this type of information. Third, we have only used one hospital dataset; multiple hospital datasets would make our model more reliable.

Future Works

This was the first part of our work. In the future, we will integrate our prediction model into the CDSS in order to check the feasibility of our model. It will help to reduce alert fatigue and result in a sophisticated CDSS by providing soft or passive alerts. Moreover, we will also try to get feedback from physicians about our prediction model (see Figure 5).

Figure 5. Future direction of this study.



Conclusions

The findings of the study showed the potential for machine learning prediction models to predict physicians' responses with high sensitivity and specificity. Among the five machine learning algorithms, the ANN model showed greater performance than the other models. This model can be a promising tool to reduce alert fatigue from CDSSs in clinical settings and can help to correctly identify an individual's alert acceptance rate.

Acknowledgments

We would like to thank AESOP (AI-Enhanced Safety of Prescription) Technology for giving us data and technological support to conduct this study. This research was funded, in part, by the Ministry of Education (MOE) (grant numbers MOE 109-6604-001-400 and DP2-109-21121-01-A-01) and the Ministry of Science and Technology (MOST) (grant number MOST109-2823-8-038-004).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of alerts in different departments and problem fitting checks.
[DOCX File , 95 KB - medinform_v8i11e19489_app1.docx]

References

1. Kruse CS, Ehrbar N. Effects of computerized decision support systems on practitioner performance and patient outcomes: Systematic review. JMIR Med Inform 2020 Aug 11;8(8):e17283 [FREE Full text] [doi: 10.2196/17283] [Medline: 32780714]
2. Campbell E, Guappone K, Sittig D, Dykstra R, Ash JS. Computerized provider order entry adoption: Implications for clinical workflow. J Gen Intern Med 2009 Jan;24(1):21-26 [FREE Full text] [doi: 10.1007/s11606-008-0857-9] [Medline: 19020942]
3. Monteiro L, Maricoto T, Solha I, Ribeiro-Vaz I, Martins C, Monteiro-Soares M. Reducing potentially inappropriate prescriptions for older patients using computerized decision support tools: Systematic review. J Med Internet Res 2019 Nov 14;21(11):e15385 [FREE Full text] [doi: 10.2196/15385] [Medline: 31724956]
4. Eiermann B, Rahmner P, Korkmaz S, Landberg C, Lilja B, Shemeikka T. Knowledge bases for clinical decision support in drug prescribing: Development, quality assurance, management, integration, implementation and evaluation of clinical value. In: Jao C, editor. Decision Support Systems. London, UK: IntechOpen; 2010.
5. Coleman JJ, van der Sijs H, Haefeli WE, Slight SP, McDowell SE, Seidling HM, et al. On the alert: Future priorities for alerts in clinical decision support for computerized physician order entry identified from a European workshop. BMC Med Inform Decis Mak 2013 Oct 01;13:111 [FREE Full text] [doi: 10.1186/1472-6947-13-111] [Medline: 24083548]
6. Powers E, Shiffman R, Melnick E, Hickner A, Sharifi M. Efficacy and unintended consequences of hard-stop alerts in electronic health record systems: A systematic review. J Am Med Inform Assoc 2018 Nov 01;25(11):1556-1566 [FREE Full text] [doi: 10.1093/jamia/ocy112] [Medline: 30239810]
7. Ko Y, Abarca J, Malone DC, Dare DC, Geraets D, Houranieh A, et al. Practitioners' views on computerized drug-drug interaction alerts in the VA system. J Am Med Inform Assoc 2007 Jan 01;14(1):56-64. [doi: 10.1197/jamia.m2224]
8. Poly TN, Islam M, Yang H, Li Y. Appropriateness of overridden alerts in computerized physician order entry: Systematic review. JMIR Med Inform 2020 Jul 20;8(7):e15653 [FREE Full text] [doi: 10.2196/15653] [Medline: 32706721]
9. Glassman PA, Simon B, Belperio P, Lanto A. Improving recognition of drug interactions: Benefits and barriers to using automated drug alerts. Med Care 2002 Dec;40(12):1161-1171. [doi: 10.1097/00005650-200212000-00004] [Medline: 12458299]
10. Tao L, Zhang C, Zeng L, Zhu S, Li N, Li W, et al. Accuracy and effects of clinical decision support systems integrated with BMJ best practice-aided diagnosis: Interrupted time series study. JMIR Med Inform 2020 Jan 20;8(1):e16912 [FREE Full text] [doi: 10.2196/16912] [Medline: 31958069]
11. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006 Mar 01;13(2):138-147. [doi: 10.1197/jamia.m1809]
12. Isaac T, Weissman JS, Davis RB, Massagli M, Cyrulik A, Sands DZ, et al. Overrides of medication alerts in ambulatory care. Arch Intern Med 2009 Feb 09;169(3):305-311. [doi: 10.1001/archinternmed.2008.551] [Medline: 19204222]
13. Nanji KC, Slight SP, Seger DL, Cho I, Fiskio JM, Redden LM, et al. Overrides of medication-related clinical decision support alerts in outpatients. J Am Med Inform Assoc 2014;21(3):487-491 [FREE Full text] [doi: 10.1136/amiajnl-2013-001813] [Medline: 24166725]
14. Topaz M, Seger DL, Slight SP, Goss F, Lai K, Wickner PG, et al. Rising drug allergy alert overrides in electronic health records: An observational retrospective study of a decade of experience. J Am Med Inform Assoc 2016 May;23(3):601-608. [doi: 10.1093/jamia/ocv143] [Medline: 26578227]
15. Getty DJ, Swets JA, Pickett RM, Gonthier D. System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. J Exp Psychol Appl 1995;1(1):19-33. [doi: 10.1037/1076-898x.1.1.19]

16. Seidling HM, Schmitt SPW, Bruckner T, Kaltschmidt J, Pruszydlo MG, Senger C, et al. Patient-specific electronic decision support reduces prescription of excessive doses. Qual Saf Health Care 2010 Oct;19(5):e15. [doi: 10.1136/qshc.2009.033175] [Medline: 20427312]
17. Del Beccaro MA, Villanueva R, Knudson KM, Harvey EM, Langle JM, Paul W. Decision support alerts for medication ordering in a computerized provider order entry (CPOE) system. Appl Clin Inform 2017 Dec 16;01(03):346-362. [doi: 10.4338/aci-2009-11-ra-0014]
18. Paterno MD, Maviglia SM, Gorman PN, Seger DL, Yoshida E, Seger AC, et al. Tiering drug-drug interaction alerts by severity increases compliance rates. J Am Med Inform Assoc 2009 Jan 01;16(1):40-46. [doi: 10.1197/jamia.m2808]
19. van der Sijs H, Aarts J, van Gelder T, Berg M, Vulto A. Turning off frequently overridden drug alerts: Limited opportunities for doing it safely. J Am Med Inform Assoc 2008;15(4):439-448 [FREE Full text] [doi: 10.1197/jamia.M2311] [Medline: 18436915]
20. Gardner RM, Evans RS. Using computer technology to detect, measure, and prevent adverse drug events. J Am Med Inform Assoc 2004 Nov 01;11(6):535-536. [doi: 10.1197/jamia.m1651]
21. Wu C, Hsu W, Islam MM, Poly TN, Yang H, Nguyen P, et al. An artificial intelligence approach to early predict non-ST-elevation myocardial infarction patients with chest pain. Comput Methods Programs Biomed 2019 May;173:109-117. [doi: 10.1016/j.cmpb.2019.01.013] [Medline: 31046985]
22. Islam MM, Poly TN, Walther BA, Yang HC, Li Y. Artificial intelligence in ophthalmology: a meta-analysis of deep learning models for retinal vessels segmentation. J Clin Med 2020 Apr 03;9(4):1018 [FREE Full text] [doi: 10.3390/jcm9041018] [Medline: 32260311]
23. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of predicting health care utilization via web search behavior: A data-driven analysis. J Med Internet Res 2016 Sep 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
24. Heiat A. Comparison of artificial neural network and regression models for estimating software development effort. Inf Softw Technol 2002 Dec;44(15):911-922. [doi: 10.1016/s0950-5849(02)00128-3]
25. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002 Dec. URL: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf [accessed 2020-10-11]
26. Zhang HJA. The optimality of naive Bayes. In: Proceedings of the Nineteenth National Conference on Artificial Intelligence. 2004 Presented at: Nineteenth National Conference on Artificial Intelligence; July 25-29, 2004; San Jose, CA.
27. Maron ME. Automatic indexing: An experimental inquiry. J ACM 1961 Jul;8(3):404-417. [doi: 10.1145/321075.321084]
28. Chen Z, Zhang T, Zhang R, Zhu Z, Yang J, Chen P, et al. Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China. Atmos Environ 2019 Apr;202:180-189. [doi: 10.1016/j.atmosenv.2019.01.027]
29. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21 [FREE Full text] [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
30. Chen T, He T. xgboost: eXtreme Gradient Boosting. The Comprehensive R Archive Network. 2020 Sep 02. URL: https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf [accessed 2020-10-11]
31. Wong HB, Lim GH. Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. Proc Singapore Healthc 2011 Dec;20(4):316-318. [doi: 10.1177/201010581102000411]
32. Abramson EL, Pfoh ER, Barrón Y, Quaresimo J, Kaushal R. The effects of electronic prescribing by community-based providers on ambulatory medication safety. Jt Comm J Qual Patient Saf 2013 Dec;39(12):545-552. [doi: 10.1016/s1553-7250(13)39070-9]
33. Graber M, Gordon R, Franklin N. Reducing diagnostic errors in medicine: What's the goal? Acad Med 2002 Oct;77(10):981-992. [doi: 10.1097/00001888-200210000-00009] [Medline: 12377672]
34. Légat L, Van Laere S, Nyssen M, Steurbaut S, Dupont AG, Cornu P. Clinical decision support systems for drug allergy checking: Systematic review. J Med Internet Res 2018 Sep 07;20(9):e258 [FREE Full text] [doi: 10.2196/jmir.8206] [Medline: 30194058]
35. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: A review. J Am Med Inform Assoc 2007 Jan 01;14(1):29-40. [doi: 10.1197/jamia.m2170]
36. Blecker S, Pandya R, Stork S, Mann D, Kuperman G, Shelley D, et al. Interruptive versus noninterruptive clinical decision support: Usability study. JMIR Hum Factors 2019 Apr 17;6(2):e12469 [FREE Full text] [doi: 10.2196/12469] [Medline: 30994460]
37. Carli D, Fahrni G, Bonnabry P, Lovis C. Quality of decision support in computerized provider order entry: Systematic literature review. JMIR Med Inform 2018 Jan 24;6(1):e3 [FREE Full text] [doi: 10.2196/medinform.7170] [Medline: 29367187]
38. Weingart SN, Toth M, Sands DZ, Aronson MD, Davis RB, Phillips RS. Physicians' decisions to override computerized drug alerts in primary care. Arch Intern Med 2003 Nov 24;163(21):2625-2631. [doi: 10.1001/archinte.163.21.2625] [Medline: 14638563]

39. Bell H, Garfield S, Khosla S, Patel C, Franklin BD. Mixed methods study of medication-related decision support alerts experienced during electronic prescribing for inpatients at an English hospital. Eur J Hosp Pharm 2019 Nov;26(6):318-322 [FREE Full text] [doi: 10.1136/ejhpharm-2017-001483] [Medline: 31798854]
40. Heringa M, Siderius H, Floor-Schreudering A, de Smet PAGM, Bouvy ML. Lower alert rates by clustering of related drug interaction alerts. J Am Med Inform Assoc 2017 Jan;24(1):54-59. [doi: 10.1093/jamia/ocw049] [Medline: 27107437]
41. Baysari MT, Tariq A, Day RO, Westbrook JI. Alert override as a habitual behavior: A new perspective on a persistent problem. J Am Med Inform Assoc 2017 Mar 01;24(2):409-412. [doi: 10.1093/jamia/ocw072] [Medline: 27274015]
42. Nanji K, Seger D, Slight S, Amato M, Beeler P, Her Q, et al. Medication-related clinical decision support alert overrides in inpatients. J Am Med Inform Assoc 2018 May 01;25(5):476-481. [doi: 10.1093/jamia/ocx115] [Medline: 29092059]
43. Yeh M, Chang Y, Wang P, Li Y, Hsu C. Physicians' responses to computerized drug-drug interaction alerts for outpatients. Comput Methods Programs Biomed 2013 Jul;111(1):17-25. [doi: 10.1016/j.cmpb.2013.02.006] [Medline: 23608682]

Abbreviations

AESOP: AI-Enhanced Safety of Prescription
ANN: artificial neural network
ATC: Anatomical Therapeutic Chemical
AUROC: area under the receiver operating characteristic curve
CDSS: clinical decision support system
CPOE: computerized physician order entry
EHR: electronic health record
GB: gradient boosting
ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification
MOE: Ministry of Education
MOST: Ministry of Science and Technology
NB: naïve Bayes
NPV: negative predictive value
PPV: positive predictive value
ReLU: rectified linear unit
RF: random forest
SVM: support vector machine

Edited by G Eysenbach, R Kukafka; submitted 20.04.20; peer-reviewed by L Zhang, T Goodwin, H Demir, S Sarbadhikari; comments to author 10.09.20; revised version received 12.09.20; accepted 19.09.20; published 19.11.20.

Please cite as:
Poly TN, Islam M, Muhtar MS, Yang HC, Nguyen PA(, Li YC(
Machine Learning Approach to Reduce Alert Fatigue Using a Disease Medication–Related Clinical Decision Support System: Model Development and Validation
JMIR Med Inform 2020;8(11):e19489
URL: https://medinform.jmir.org/2020/11/e19489
doi:10.2196/19489
PMID:33211018

Original Paper

Deep Learning–Based Detection of Early Renal Function Impairment Using Retinal Fundus Images: Model Development and Validation

Eugene Yu-Chuan Kang[1,2], MD; Yi-Ting Hsieh[3], MD, PhD; Chien-Hung Li[4], MSc; Yi-Jin Huang[4], MSc; Chang-Fu Kuo[2,5], MD, PhD; Je-Ho Kang[6], MD; Kuan-Jen Chen[1,2], MD; Chi-Chun Lai[1,2], MD; Wei-Chi Wu[1,2], MD, PhD; Yih-Shiou Hwang[1,2], MD, PhD

[1]Department of Ophthalmology, Chang Gung Memorial Hospital, Linkou Medical Center, Taoyuan, Taiwan

[2]College of Medicine, Chang Gung University, Taoyuan, Taiwan

[3]Department of Ophthalmology, National Taiwan University Hospital, Taipei, Taiwan

[4]Acer Healthcare Incorporated, New Taipei, Taiwan

[5]Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Linkou Medical Center, Taoyuan, Taiwan

[6]Department of Nephrology, Yang Ming Hospital, Taoyuan, Taiwan

Corresponding Author:
Yih-Shiou Hwang, MD, PhD
Department of Ophthalmology
Chang Gung Memorial Hospital, Linkou Medical Center
No. 5, Fu-Hsin Rd.
Taoyuan, 333
Taiwan
Phone: 886 3 3281200 ext 8666
Fax: 886 3 3287798
Email: yihshiou.hwang@gmail.com

Abstract

Background: Retinal imaging has been applied for detecting eye diseases and cardiovascular risks using deep learning–based methods. Furthermore, retinal microvascular and structural changes were found in renal function impairments. However, a deep learning–based method using retinal images for detecting early renal function impairment has not yet been well studied.

Objective: This study aimed to develop and evaluate a deep learning model for detecting early renal function impairment using retinal fundus images.

Methods: This retrospective study enrolled patients who underwent renal function tests with color fundus images captured at any time between January 1, 2001, and August 31, 2019. A deep learning model was constructed to detect impaired renal function from the images. Early renal function impairment was defined as estimated glomerular filtration rate <90 mL/min/1.73 m^2. Model performance was evaluated with respect to the receiver operating characteristic curve and area under the curve (AUC).

Results: In total, 25,706 retinal fundus images were obtained from 6212 patients for the study period. The images were divided at an 8:1:1 ratio. The training, validation, and testing data sets respectively contained 20,787, 2189, and 2730 images from 4970, 621, and 621 patients. There were 10,686 and 15,020 images determined to indicate normal and impaired renal function, respectively. The AUC of the model was 0.81 in the overall population. In subgroups stratified by serum hemoglobin A_{1c} (HbA_{1c}) level, the AUCs were 0.81, 0.84, 0.85, and 0.87 for the HbA_{1c} levels of ≤6.5%, >6.5%, >7.5%, and >10%, respectively.

Conclusions: The deep learning model in this study enables the detection of early renal function impairment using retinal fundus images. The model was more accurate for patients with elevated serum HbA_{1c} levels.

(JMIR Med Inform 2020;8(11):e23472) doi:10.2196/23472

KEYWORDS

deep learning; renal function; retinal fundus image; diabetes; renal; kidney; retinal; eye; imaging; impairment; detection; development; validation; model

Introduction

Background

Chronic kidney disease (CKD) is defined as a gradual loss of renal function, and it can progress to an advanced stage, termed end-stage renal disease (ESRD). According to the 2016 annual report of the US Renal Data System [1], the incidence of treated ESRD increased gradually at the rate of 2%-4% from 2003 to 2016 in almost one-third of all countries [1]. Taiwan, in particular, had the highest incidence of treated ESRD (493 patients per million in the general population) and the highest prevalence of treated ESRD (3392 patients per million in the general population) among all countries worldwide [1]. According to Taiwan's National Health Insurance 2018 report [2], CKD incurred the highest medical costs in the country, approximately US $1.7 billion. Therefore, progress is required in the prevention and screening of kidney disease in Taiwan. In all the etiologies of CKD, diabetes is a leading cause; it has been estimated that 1 in 4 adults with diabetes have impaired renal function [3]. Therefore, the monitoring of renal function is especially important for patients with diabetes; it is also crucial in countries where ESRD is prevalent.

With the increasing sophistication of artificial intelligence, deep learning has been increasingly applied to various types of medical imaging analysis, especially ophthalmology imaging [4]. Among ophthalmology imaging techniques, retinal imaging has been used to establish deep learning models for detecting not only eye diseases (eg, diabetic retinopathy and glaucoma) [5,6] but also systemic cardiovascular risks [7]. The microvascular network in the retina can be easily observed; it is structurally and physiologically similar to the vascular structures of many other systems or organs and can be used in the evaluation of various disorders, including systemic hypertension, coronary artery disease, and central nervous disorders [8-10]. Studies have also demonstrated that changes in the retinal vasculature are associated with renal dysfunction and reduced estimated glomerular filtration rate (eGFR) [11,12].

Objective

Scholars have recommended applying artificial intelligence to the management and prevention of kidney disease [13]. However, few studies have developed deep learning–based methods for detecting early renal function impairment from retinal images. Therefore, we established a deep learning model to detect early renal function impairment from retinal fundus images. We also evaluated the performance of our model when applied to patients with diabetes.

Methods

Study Population

In this retrospective study, we included patients who underwent retinal fundus imaging examinations and laboratory tests at any time between January 1, 2001, and August 31, 2019, at Chang Gung Memorial Hospital (CGMH), Linkou Medical Center, Taoyuan, Taiwan. The retinal fundus images were taken with fundus cameras (Topcon Medical Systems, KOWA, and Digital Non-Mydriatic Retinal Camera, Canon). The laboratory tests conducted for serum creatinine and serum hemoglobin A_{1c} (HbA_{1c}) were respectively performed with a colorimetric method and high-performance liquid chromatography at the CGMH Department of Laboratory Medicine. Demographic data, including those on age and sex, were also retrieved from CGMH's electronic medical record system. This study was approved by the CGMH Institutional Review Board (CGMH IRB No. 201901544B0), and the requirement for informed consent was waived because patient data were deidentified. The study was conducted in accordance with the Declaration of Helsinki.

Data Management

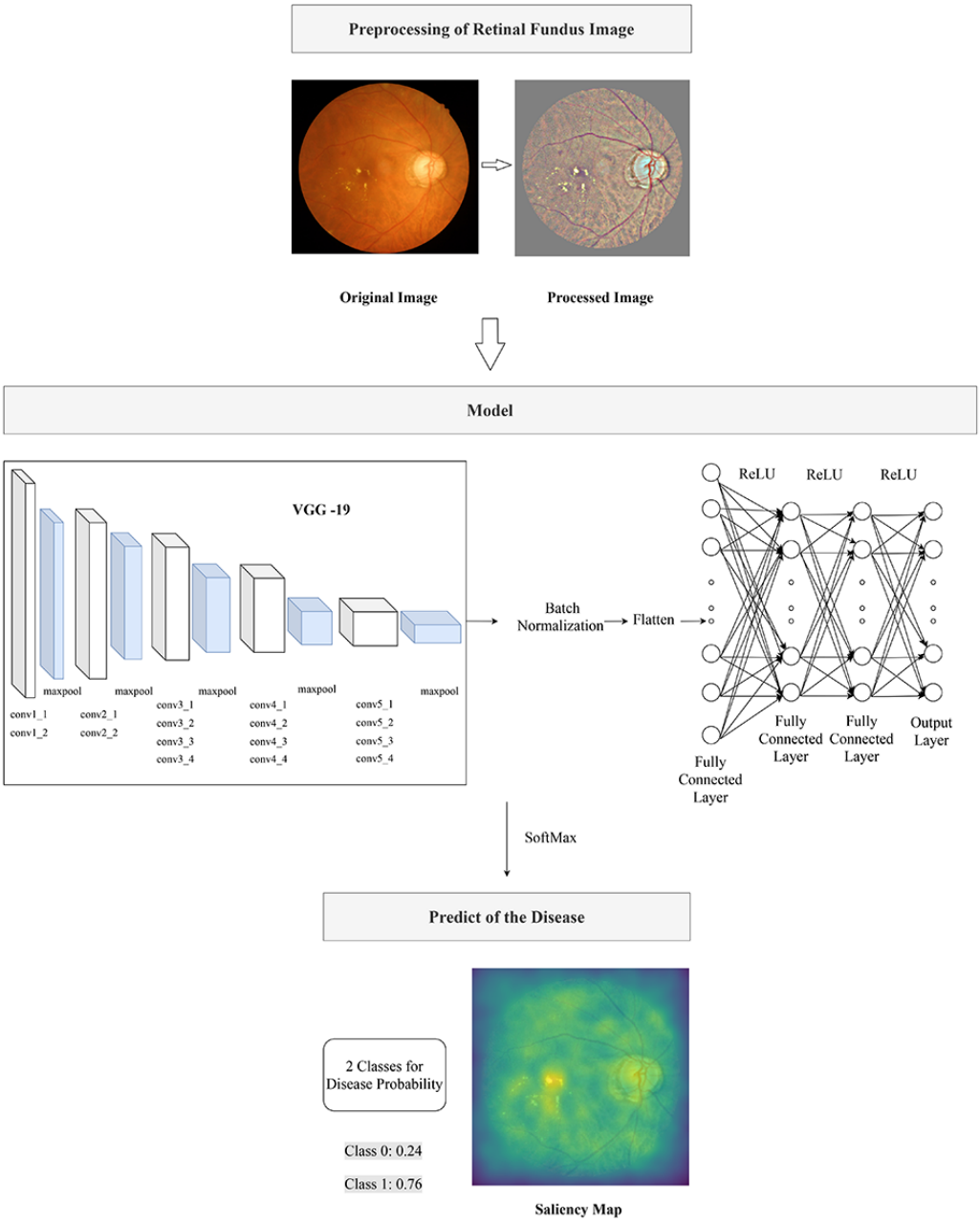
After the data were retrieved, retinal fundus images were linked to the corresponding renal functions, which were measured by eGFR. In our study, the eGFR was calculated using the Modification of Diet in Renal Disease (MDRD) equation, which includes the patient's age, sex, and serum creatinine, as revised by Levey et al [14]. We defined early renal function impairment as eGFR <90 mL/min/1.73 m^2, which was equal to or more severe than the mildly decreased glomerular filtration rate according to the definition published in the 2012 guidelines of "Kidney Disease: Improving Global Outcomes" [15]. We only included laboratory tests that had been conducted within 3 months before or after the corresponding retinal fundus images were captured. Patients without available serum creatinine results were excluded. We deidentified the data after the images and laboratory data were linked. Subsequently, we excluded retinal fundus images that had color filters, were merged, or were neither macula- nor disk-centered. For an image to be included, both the macula and disk were required to be visible. We also excluded poor-quality images, such as those that had a low resolution, were out of focus, had a large halo, or had a large shadow. Multimedia Appendix 1 presents some examples of the excluded images.

Model Architecture

The model architecture is illustrated in Figure 1. To reduce the variation of illumination and camera resolution between the different retinal images, all images were processed using the method proposed by Graham [16]. All images were resized to a resolution of 224 × 224 × 3 and were processed to reduce variance in illumination between images before running the algorithm. For the convolutional neural network (CNN), we selected VGG-19 formulated by the Visual Geometry Group [17]. We selected VGG-19 because it exhibited the best performance in our preliminary model training relative to ResNet, Inception V3, and Inception V4. Furthermore, in previous research, VGG-19 exhibited comparable performance to other deeper CNNs in medical imaging analysis in general and in ophthalmological imaging in particular [18]. After the CNN retrieved the image feature, a batch normalization layer was added to accelerate training, and the features were flattened to 1-dimensional vectors. Subsequently, we added 3 fully connected layers that had a nonlinear rectified linear unit (ReLU) activation function and 1 final output layer with the softmax activation function. The results were classified results into 2 classes—class 0 and 1, which represented normal and impaired renal function, respectively. The probability for each class was

presented. As presented in Figure 1, the probability of disease was 0.76, and a saliency map was generated based on the features marked as determinative for the detection of renal function impairment.

Figure 1. Architecture of the model for detecting early renal function impairment from retinal fundus images. ReLU: rectified linear unit.



Model Training and Performance

The data sets of all patients were partitioned into nonoverlapping training, validation, and testing sets at an 8:1:1 ratio, and the images from each patient were linked to the corresponding renal function results. The model was trained, validated, and tested on the basis of the images. The model was trained on a workstation with an Intel Xeon Silver 4110 CPU at 2.10 GHz, a NVIDIA GeForce GTX 1080 Ti (with 11 GB of video memory) graphics card, and 125 GB of RAM. For this model, the learning rate and batch size were set as 0.000005 and 32, respectively. An Adam optimizer was used, and the model was trained up to 120 epochs. The model was established based on the achievement of maximum accuracy and minimum loss in the validation set. The learning curve of the model is presented in Multimedia Appendix 2. To analyze the model prediction, we generated saliency maps (Figure 1), which identified the region of the retinal fundus photo that contributed to the model's determination of renal function impairment. We also classified the testing set according to the patient's HbA_{1c} levels. Furthermore, the model performance was evaluated at HbA_{1c} levels of ≤6.5%, >6.5%, >7.5%, and >10.0% in the testing data set.

Statistical Analysis

For the demographic data, continuous variables were expressed in terms of the mean (SD). Chi-square tests and *t* tests were conducted for descriptive analyses of categorical (sex) and continuous (age and HbA_{1c}) variables, respectively. To analyze

the performance of our model, receiver operating characteristic (ROC) curves were plotted, and the area under the curve (AUC) for each ROC curve was calculated. AUC values of 0.7-0.8 and >0.8 indicated acceptable discrimination and excellent discrimination, respectively. An AUC value of 1 represented perfect discrimination, and AUC value of 0.5 represented no or random discrimination [19]. We also measured the sensitivity, specificity, positive predictive value (PPV), and accuracy of the model. Model performance was evaluated using the images in the testing set. Statistical significance was indicated if $P<.05$. Statistical analyses were conducted using SPSS (Version 23, IBM Corp).

Results

Demographic Characteristics

In this study, we initially included 7167 patients with 51,666 retinal fundus images. We then excluded 13.32% (955/7167) patients and 50.24% (25,960/51,666) images after applying the exclusion criteria. The remaining 25,706 retinal fundus images from 6212 patients were included in the final analysis, and each patient may have a different number of images. The variance was 1 to 33 images per patient. The training, validation, and testing sets comprised 20,787, 2189, and 2730 images from 4970, 621, and 621 patients, respectively (Table 1).

Table 1. Distribution of patients with clinical information in the training, validation, and testing groups.

Characteristic	Total (N=6212)	Training (n=4970)	Validation (n=621)	Testing (n=621)
Sex, n (%)				
Male	3363 (54.14)	2689 (54.10)	339 (54.6)	335 (53.9)
Female	2849 (45.86)	2281 (45.90)	282 (45.4)	286 (46.1)
Age (years), mean (SD)	57.6 (16.6)	58.7 (15.9)	51.0 (19.1)	51.6 (17.4)
eGFR[a] (ml/min/1.73 m^2), mean (SD)	78.6 (32.6)	77.8 (32.2)	86.5 (34.1)	80.4 (35.6)
HbA_{1c}[b] (%), mean (SD)	7.6 (2.0)	7.6 (1.9)	7.6 (1.8)	7.9 (2.1)

[a]eGFR: estimated glomerular filtration rate.

[b]HbA_{1c}: hemoglobin A_{1c}.

Each patient was randomly assigned to a group, and all images from a patient belonged only to the group the patient was assigned to. With regard to demographic characteristics, 54.14% (3363/6212) of the patients were male, and the mean age of all patients was 57.6 (SD 16.6) years. As for clinical characteristics, the mean eGFR and serum HbA_{1c} levels were 78.6 mL/min/1.73 m^2 (SD 32.6) and 7.6% (SD 2.0%), respectively. Table 2 presents the clinical information for normal and impaired renal function (eGFR <90 mL/min/1.73 m^2) in our study population.

Table 2. Clinical information of patients with normal or impaired renal function (N=6212); all *P* values are <.001.

Characteristic	Normal renal function (n=3108)	Impaired renal function (n=3104)
Sex, n (%)		
Male	1539 (49.52)	1824 (58.76)
Female	1569 (50.48)	1280 (41.24)
Age (years), mean (SD)	47.2 (16.1)	64.1 (13.1)
HbA_{1c}[a] (%), mean (SD)	7.7 (2.1)	7.5 (1.9)

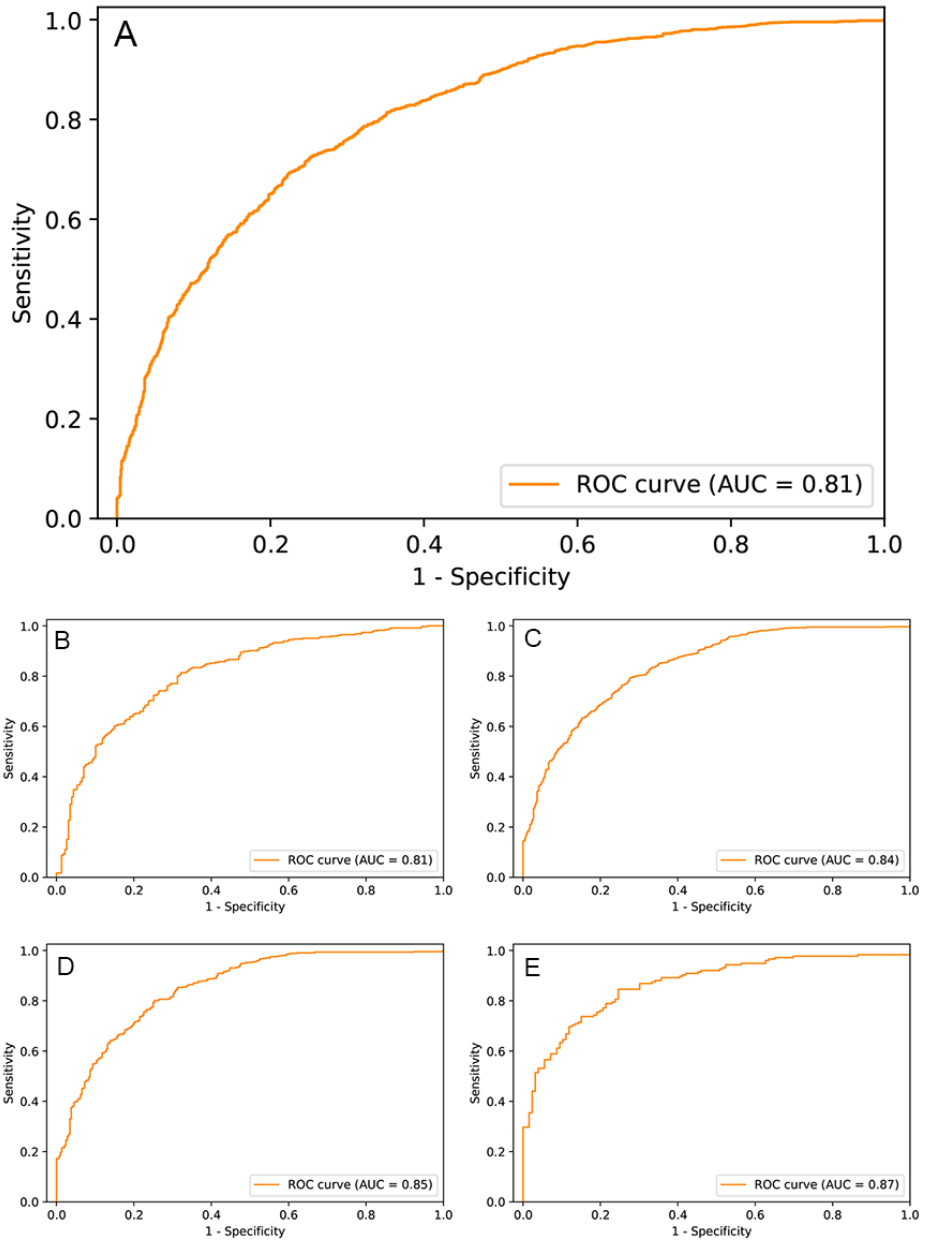
[a]HbA_{1c}: hemoglobin A_{1c}.

Compared with patients with healthy renal function, patients with impaired renal function were more likely to be male (impaired vs healthy: 58.3% vs 49.1%; $P<.001$), older adults (64.1 years vs 47.2 years; $P<.001$), and with a lower serum HbA_{1c} level (7.5% vs 7.7%; $P<.001$). Multimedia Appendix 3 shows the clinical information in patients with stratified HbA_{1c} levels in the testing set.

Model Performance

The ROC curves obtained from tests of our model are presented in Figure 2. Model performance for subgroups stratified by serum HbA_{1c} level was also tested. Model performance increased gradually with serum HbA_{1c} level.

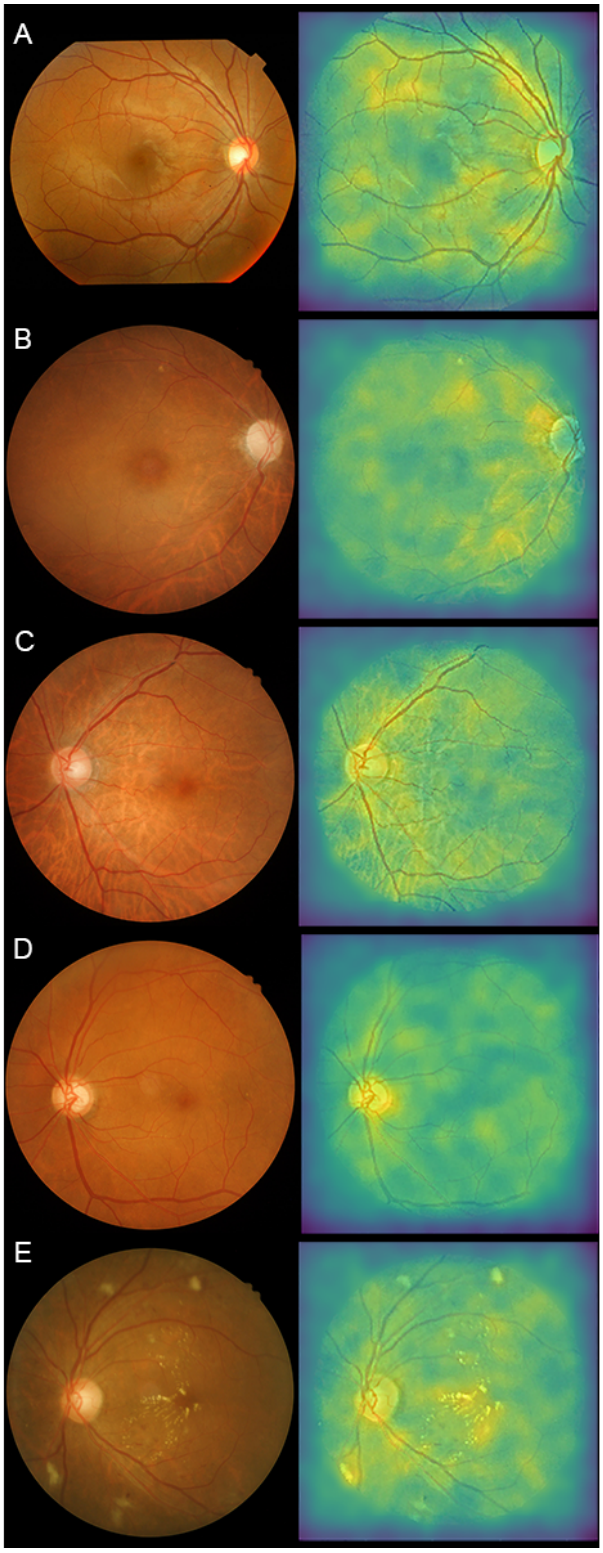
Figure 2. ROC curves for the model in detecting early renal function impairment in different groups of patients. ROC curves for (A) all patients (AUC = 0.81, sensitivity = 0.83, specificity = 0.62, PPV = 0.73, accuracy = 0.73); (B) patients with $HbA_{1c} \leq 6.5\%$ (AUC = 0.81, sensitivity = 0.84, specificity = 0.62, PPV = 0.77, accuracy = 0.75), (C) patients with $HbA_{1c} > 6.5\%$ (AUC = 0.84, sensitivity = 0.89, specificity = 0.61, PPV = 0.77, accuracy = 0.77), (D) patients with $HbA_{1c} > 7.5\%$ (AUC = 0.85, sensitivity = 0.89, specificity = 0.60, PPV = 0.82, accuracy = 0.79), and (E) patients with $HbA_{1c} > 10.0\%$ (AUC = 0.87, sensitivity = 0.89, specificity = 0.61, PPV = 0.77, accuracy = 0.77). AUC: area under the curve; HbA_{1c}: hemoglobin A_{1c}; PPV: positive predictive value; ROC: receiver operating characteristic.



Saliency Maps

Representative saliency maps are presented in Figure 3, where the regions responsible for the prediction of impaired renal function are highlighted in the lighter color. In Figure 3, the retinal-vessel features are marked for a true-positive case with a relatively normal retinal fundus image. Common signs of retina abnormality, such as exudation, hemorrhage, and drusen, also played a role in the detection of renal function impairment.

Figure 3. Selected retinal fundus images and their corresponding saliency maps in true-negative and true-positive cases. (A) No renal function impairment detected. Patient's eGFR = 102.6 mL/min/1.73 m^2 and HbA_{1c} = 13.4%. (B) Renal function impairment detected. Patient's eGFR = 40.0 mL/min/1.73 m^2 and HbA_{1c} = 5.1%. (C) Renal function impairment detected. Patient's eGFR = 50 mL/min/1.73 m^2 and HbA_{1c} = 6.5%. (D) Renal function impairment detected. Patient's eGFR = 80.5 mL/min/1.73 m^2 and HbA_{1c} = 7.3%. (E) Renal function impairment detected. Patient's eGFR = 67.7 ml/min/1.73 m^2 and HbA_{1c} = 8.9%. eGFR: estimated glomerular filtration rate; HbA_{1c}: hemoglobin A_{1c}.



Discussion

Main Findings

In this study, we developed a deep learning model for detecting early renal function impairment from retinal fundus images. The AUC of the model was 0.81 for the detection of early renal function impairment in the general population, and the model performed better when applied to patients with diabetes or patients with elevated serum HbA_{1c} levels.

Importance of Renal Function Screening

The 2016 annual report of the US Renal Data System [1] notes that ESRD is becoming increasingly prevalent in many countries, underscoring the increased burdens of CKD and ESRD on society. Taiwan has a high incidence and prevalence of CKD and ESRD, and the country bears significant health care burden associated with CKD and ESRD [1,2]. Thus, several studies in Taiwan have evaluated the etiology and screening of kidney diseases [20,21]. In Taiwan, CKD prevention has been hampered by low public awareness, infrequent eGFR measurements, and delayed referrals [22,23]. Although a study suggested the importance of comprehensive renal function screening in high-risk populations, such as patients with diabetes [15], evidence for the cost-effectiveness and benefits of routine screening for CKD remain inconclusive because commonly used tests with urine or blood are inconvenient and invasive [24].

Deep Learning in Renal Function Using Ultrasonography

Deep learning methods provide a potential solution to this problem. With the increasing sophistication of artificial intelligence, deep learning has been increasingly applied in various fields, including medicine [25]. The use of artificial intelligence for management of kidney disease has been recently proposed, and its potential has been well recognized by physicians [13]. Kuo et al [26] developed a deep learning model for predicting renal function by using kidney ultrasound images. Their model was more accurate (0.86) in detecting cases with eGFR <60 mL/min/1.73 m^2 than the judgments of experienced nephrologists (0.60-0.80). Although our model had lower overall accuracy (0.73 for all patients and 0.79 for patients with HbA_{1c} > 7.5%) relative to theirs, our model's accuracy is still comparable with that of the judgments of experienced nephrologists employing ultrasound images. Moreover, our model could detect early renal function impairment with eGFR <90 mL/min/1.73 m^2, a functionality that was not evaluated by Kuo et al [26].

Deep Learning Using Retinal Fundus Images

Retinal fundus imaging can be executed even by untrained medical staff and has high accessibility. Furthermore, a patient's retinal fundus images can be captured in less than 10 minutes, and the patient can be promptly referred to a specialist if a problem is detected [27]. A previous review on deep learning in ophthalmology noted that retinal fundus images can be used to identify several eye diseases, including glaucoma, macular degeneration, refractive errors, and, most importantly, diabetic retinopathy [18]. Furthermore, systemic cardiovascular risks can also be determined from retinal images [7]. Those results suggest the potential of using retinal photography for large-scale disease screening.

Using Retinal Fundus Images for Renal Function Prediction

In our study, we developed a deep learning model to detect early renal function impairment. The model had excellent discrimination (AUC=0.81; excellent discrimination was defined as AUC >0.8) [19]. The saliency maps revealed that features in retinal vasculature and of hemorrhages and exudations were influential in the determination of impaired renal function. This finding is compatible with the findings of previous reports on specific retinal microvascular and structural changes in renal function impairment [11,28]. When applied to patients with diabetes, our model had a sensitivity as high as 0.89 but a specificity of only 0.60. We noted that our model produced several false positives for patients who shared some similar ophthalmic pathologies presenting on the fundus images. These pathologies included subretinal fluid, optic disc swelling caused by optic neuritis, and retinal scarring (Multimedia Appendix 4). However, no robust association between these pathologies and renal function is indicated in the literature. As noted in the saliency maps, the model identified retinal vessel characteristics and the presence of hemorrhage and exudation. Subretinal fluid and optic disc swelling may alter retinal vascular features and thus affect the model prediction. Ocular infection or inflammation was also presented with retinal vascular change, hemorrhage, exudation, and pigmented scars [29], which may be similar to the retinal presentation of impaired renal function. Therefore, these coexisting ocular pathologies may have reduced model specificity. For future studies on deep learning, we suggest the use of multimodal retinal images to predict renal function impairment; the analysis of multimodal retinal images has been reported to yield greater accuracy in diagnosing age-related macular degeneration [30].

Comparison of Model Performance in Diabetes and Between the Previous Study

Our model had a greater AUC and sensitivity for higher HbA_{1c} levels (up to AUC=0.87 for HbA_{1c} >10%). Some possible explanations for this performance include more profound microvascular damage in patients with worse glucose control and the coexistence of signs of diabetic retinopathy and diabetic nephropathy, which were noted to be significantly associated [31,32]. A deep learning algorithm was recently formulated by a research group at the Singapore National Eye Center (SNEC) [33]. Their algorithm was used to detect CKD with eGFR <60 mL/min/1.73 m^2 by using both retinal images and risk factors, individually and in combination, in 3 population-based screening databases from Singapore and China [33]. Their image-based model had an AUC of 0.91 in their internal validation (Singapore Epidemiology of Eye Diseases database), AUCs of 0.73 and 0.84 in their external testing (Singapore Prospective Study Program and Beijing Eye Study, respectively), and an AUC of 0.89 when applied in patients with diabetes. The overall performance of our model (AUC=0.81) is in between the performance levels of their model in their internal validation

and external testing. This difference in performance is attributable to differences in patient characteristics or model architecture. Our hospital is a referral medical center with comprehensive ophthalmology equipment for the management of advanced eye diseases [34]. Compared with population-based screening databases, our database featured more patients with pathologies on the retina or other parts of the eye, which may have increased the likelihood of model misdiagnosis [35]. In addition, the model was trained using images from 1 of 3 types of fundus cameras and 1 of 2 different image formats (JPEG or PNG). This variety likely affected the predictive performance of the model. Specifically, when our model was applied to the subgroup of patients with diabetes, its performance (AUC = 0.84 in HbA_{1c} >6.5%, 0.85 in HbA_{1c} >7.5%, and 0.87 in HbA_{1c} >10.0%) was comparable to that of the SNEC model.

Study Limitations

Our study has some limitations. First, the results of the MDRD formula for calculating eGFR did not reflect definite renal function; variations related to ethnicity have been reported, and this measure was noted to be less accurate when applied to the Taiwanese population [21,36]. Second, as we aimed to detect early renal function impairment (ie, eGFR <90 mL/min/1.73 m^2), we did not test the efficacy of our model in predicting advanced kidney diseases. Third, we discarded poor-quality fundus images before training the model. However, poor-quality images are encountered in clinical settings, and model performance may thus be affected by factors such as patient cooperation and medial opacities of the eye and small pupils [27]. Although retinal fundus imaging is a relatively accessible test, the feasibility of our model in real-world applications requires further investigation. Fourth, the model's detection of renal function may be affected by signs from some ocular diseases that are related to neither systemic vascular function nor renal function. For example, certain retinal infections may alter the model's prediction of renal function impairment; such infections are not associated with systemic vascular function but share a common feature, namely the presence of hemorrhages or exudates on the retina. By contrast, renal function impairment with nonvascular causes, such as urinary tract obstruction, may not present vasculature or retinal abnormality in fundus images during the early disease phase. In our study, selection bias may have occurred in the subpopulation with a referral medical center. This subpopulation has a higher proportion of patients with ocular diseases coexisting with other organic diseases. Fifth, we did not perform patient-matching between the training, validation, and testing groups. Thus, differences in clinical characteristics may have affected the learning and performance of the model. Finally, the function of this model lies in screening rather than diagnosis. A thorough kidney examination that includes ultrasonography and insulin clearance remains crucial.

Conclusion

In conclusion, our study formulated and evaluated a deep learning model for predicting early renal function impairment. Our model also performed better, as indicated by the increased AUC, when applied to patients with diabetes or patients with elevated serum HbA_{1c} levels. Color fundus images are easy to obtain and can thus be feasibly applied to the detection of early renal function impairment, especially in patients with diabetes, in conjunction with our model.

Acknowledgments

The authors thank Acer Healthcare, Taiwan for providing technical support. The authors also thank Miranda Chun-Ya Kang and Wallace Academic Editing for English editing. This study was funded by research grants from the National Science Council, Taiwan (MOST 105-2314-B-182A-076 and MOST 106-2314-B-182A-045 -MY3), and Chang Gung Memorial Hospital, Taiwan (CMRPG3C0171). The founding organizations had no role in the interpretation of the study results.

Authors' Contributions

EYCK, CCL, WCW, and YSH contributed to the conception and design of the study. Data were collected by CFK, KJC, CCL, WCW, and YSH. Data analysis was conducted by CHL and YJH. YTH, JHK, WCW, and YSH contributed to data interpretation. EYCK wrote the manuscript.

Conflicts of Interest

CHL and YJH are employees of Acer Healthcare, Taiwan.

Multimedia Appendix 1
Excluded retinal fundus images due to (a) the use of a color filter, (b) the image not being centered on the macula or disc, (c) blurriness, (d) the presence of glares, (e) poor light exposure, and (f) an invisible macula.
[PNG File , 718 KB - medinform_v8i11e23472_app1.png]

Multimedia Appendix 2
Learning curves of the model, demonstrating the learning rate in the epochs versus (a) train loss and (b) accuracy. The epoch of 56 with the lowest validation loss was selected for testing.
[PNG File , 156 KB - medinform_v8i11e23472_app2.png]

Multimedia Appendix 3
Clinical information of patients with normal or impaired renal function stratified by different HbA1c levels in the testing set.
[DOCX File , 16 KB - medinform_v8i11e23472_app3.docx]

Multimedia Appendix 4
Retinal fundus images from false-positive cases showing (a) a swollen optic disc due to idiopathic optic neuritis, (b) a chorioretinal scar caused by previous inflammation, and (c) subretinal fluid caused by retinal detachment.
[PNG File , 408 KB - medinform_v8i11e23472_app4.png]

References

1. Saran R, Robinson B, Abbott KC, Agodoa LYC, Albertus P, Ayanian J, et al. US Renal Data System 2016 Annual Data Report: Epidemiology of Kidney Disease in the United States. Am J Kidney Dis 2017 Mar;69(3 Suppl 1):A7-A8 [FREE Full text] [doi: 10.1053/j.ajkd.2016.12.004] [Medline: 28236831]
2. 2018 National Health Insurance Medical Cost Top 20. Taiwan National Health Insurance Administration, Ministry of Health and Welfare. URL: https://www.nhi.gov.tw/Content_List.aspx?n=D529CAC4D8F8E77B&topn=23C660CAACAA159D [accessed 2020-08-02]
3. Afkarian M, Zelnick LR, Hall YN, Heagerty PJ, Tuttle K, Weiss NS, et al. Clinical Manifestations of Kidney Disease Among US Adults With Diabetes, 1988-2014. JAMA 2016 Aug 09;316(6):602-610. [doi: 10.1001/jama.2016.10924] [Medline: 27532915]
4. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
5. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 2017 Dec 12;318(22):2211-2223 [FREE Full text] [doi: 10.1001/jama.2017.18152] [Medline: 29234807]
6. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
7. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
8. London A, Benhar I, Schwartz M. The retina as a window to the brain-from eye research to CNS disorders. Nat Rev Neurol 2013 Jan;9(1):44-53. [doi: 10.1038/nrneurol.2012.227] [Medline: 23165340]
9. Tapp RJ, Owen CG, Barman SA, Welikala RA, Foster PJ, Whincup PH, et al. Associations of Retinal Microvascular Diameters and Tortuosity With Blood Pressure and Arterial Stiffness: United Kingdom Biobank. Hypertension 2019 Dec;74(6):1383-1390 [FREE Full text] [doi: 10.1161/HYPERTENSIONAHA.119.13752] [Medline: 31661987]
10. Phan K, Mitchell P, Liew G, Plant AJ, Wang SB, Xu J, et al. Severity of coronary artery disease and retinal microvascular signs in patients with diagnosed versus undiagnosed diabetes: cross-sectional study. J Thorac Dis 2016 Jul;8(7):1532-1539 [FREE Full text] [doi: 10.21037/jtd.2016.05.61] [Medline: 27499940]
11. Lim LS, Cheung CY, Sabanayagam C, Lim SC, Tai ES, Huang L, et al. Structural changes in the retinal microvasculature and renal function. Invest Ophthalmol Vis Sci 2013 Apr 26;54(4):2970-2976. [doi: 10.1167/iovs.13-11941] [Medline: 23572105]
12. Ooi QL, Tow FKNH, Deva R, Alias MA, Kawasaki R, Wong TY, et al. The microvasculature in chronic kidney disease. Clin J Am Soc Nephrol 2011 Aug;6(8):1872-1878 [FREE Full text] [doi: 10.2215/CJN.10291110] [Medline: 21784828]
13. Yuan Q, Zhang H, Deng T, Tang S, Yuan X, Tang W, et al. Role of Artificial Intelligence in Kidney Disease. Int J Med Sci 2020;17(7):970-984 [FREE Full text] [doi: 10.7150/ijms.42078] [Medline: 32308551]
14. Levey AS, Coresh J, Greene T, Stevens LA, Zhang YL, Hendriksen S, Chronic Kidney Disease Epidemiology Collaboration. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. Ann Intern Med 2006 Aug 15;145(4):247-254. [doi: 10.7326/0003-4819-145-4-200608150-00004] [Medline: 16908915]
15. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney International Supplements. URL: https://kdigo.org/wp-content/uploads/2017/02/KDIGO_2012_CKD_GL.pdf [accessed 2020-08-02]
16. Graham B. Kaggle Diabetic Retinopathy Detection Competition Report. 2015. URL: http://scholar.google.com.tw/scholar_url?url=https://kaggle-forum-message-attachments.storage.googleapis.com/88655/2795/competitionreport.pdf&hl=zh-TW&sa=X&scisig=AAGBfm0vGsKIa3EEhTDn8VZSmliV6_3Hsg&nossl=1&oi=scholarr [accessed 2020-08-02]
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv Preprint posted online on April 10, 2015 [FREE Full text]

18. Ting DSW, Peng L, Varadarajan AV, Keane PA, Burlina PM, Chiang MF, et al. Deep learning in ophthalmology: The technical and clinical considerations. Prog Retin Eye Res 2019 Sep;72:100759. [doi: 10.1016/j.preteyeres.2019.04.003] [Medline: 31048019]
19. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010 Sep;5(9):1315-1316 [FREE Full text] [doi: 10.1097/JTO.0b013e3181ec173d] [Medline: 20736804]
20. Tsai M, Hsu C, Lin M, Yen M, Chen H, Chiu Y, et al. Incidence, Prevalence, and Duration of Chronic Kidney Disease in Taiwan: Results from a Community-Based Screening Program of 106,094 Individuals. Nephron 2018;140(3):175-184. [doi: 10.1159/000491708] [Medline: 30138926]
21. Hwang S, Tsai J, Chen H. Epidemiology, impact and preventive care of chronic kidney disease in Taiwan. Nephrology (Carlton) 2010 Jun;15 Suppl 2:3-9. [doi: 10.1111/j.1440-1797.2010.01304.x] [Medline: 20586940]
22. Wen CP, Cheng TYD, Tsai MK, Chang YC, Chan HT, Tsai SP, et al. All-cause mortality attributable to chronic kidney disease: a prospective cohort study based on 462 293 adults in Taiwan. Lancet 2008 Jun 28;371(9631):2173-2182. [doi: 10.1016/S0140-6736(08)60952-6] [Medline: 18586172]
23. Lin M, Lee CT, Kuo M, Hwang S, Chen H, Chiu Y. Effects of physician's specialty on regular chronic kidney disease care in predialysis: A population-based cross-sectional study. Medicine (Baltimore) 2018 Jun;97(26):e11317 [FREE Full text] [doi: 10.1097/MD.0000000000011317] [Medline: 29953019]
24. Moyer VA, U.S. Preventive Services Task Force. Screening for chronic kidney disease: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2012 Oct 16;157(8):567-570 [FREE Full text] [doi: 10.7326/0003-4819-157-8-201210160-00533] [Medline: 22928170]
25. Amisha, Malik P, Pathania M, Rathaur VK. 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]
26. Kuo C, Chang C, Liu K, Lin W, Chiang H, Chung C, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. NPJ Digit Med 2019;2:29 [FREE Full text] [doi: 10.1038/s41746-019-0104-2] [Medline: 31304376]
27. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. 2020 Presented at: CHI Conference on Human Factors in Computing Systems (CHI '20); April 25-30, 2020; Honolulu, HI URL: https://doi.org/10.1145/3313831.3376718 [doi: 10.1145/3313831.3376718]
28. Deva R, Alias MA, Colville D, Tow FKNH, Ooi QL, Chew S, et al. Vision-threatening retinal abnormalities in chronic kidney disease stages 3 to 5. Clin J Am Soc Nephrol 2011 Aug;6(8):1866-1871 [FREE Full text] [doi: 10.2215/CJN.10321110] [Medline: 21784818]
29. Sudharshan S, Ganesh SK, Biswas J. Current approach in the diagnosis and management of posterior uveitis. Indian J Ophthalmol 2010;58(1):29-43 [FREE Full text] [doi: 10.4103/0301-4738.58470] [Medline: 20029144]
30. Vaghefi E, Hill S, Kersten HM, Squirrell D. Multimodal Retinal Image Analysis via Deep Learning for the Diagnosis of Intermediate Dry Age-Related Macular Degeneration: A Feasibility Study. J Ophthalmol 2020;2020:7493419 [FREE Full text] [doi: 10.1155/2020/7493419] [Medline: 32411434]
31. Klein R, Zinman B, Gardiner R, Suissa S, Donnelly SM, Sinaiko AR, Renin-Angiotensin System Study. The relationship of diabetic retinopathy to preclinical diabetic glomerulopathy lesions in type 1 diabetic patients: the Renin-Angiotensin System Study. Diabetes 2005 Feb;54(2):527-533 [FREE Full text] [doi: 10.2337/diabetes.54.2.527] [Medline: 15677511]
32. Lee WJ, Sobrin L, Lee MJ, Kang MH, Seong M, Cho H. The relationship between diabetic retinopathy and diabetic nephropathy in a population-based study in Korea (KNHANES V-2, 3). Invest Ophthalmol Vis Sci 2014 Sep 09;55(10):6547-6553. [doi: 10.1167/iovs.14-15001] [Medline: 25205863]
33. Sabanayagam C, Xu D, Ting D, Nusinovici S, Banu R, Hamzah H, et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. The Lancet Digital Health 2020 Jun;2(6):e295-e302 [FREE Full text] [doi: 10.1016/s2589-7500(20)30063-7]
34. Kang EY, Tai W, Lin J, Huang C, Yeh P, Wu W, et al. Eye-related Emergency Department Visits with Ophthalmology Consultation in Taiwan: Visual Acuity as an Indicator of Ocular Emergency. Sci Rep 2020 Jan 22;10(1):982 [FREE Full text] [doi: 10.1038/s41598-020-57804-2] [Medline: 31969635]
35. Hsieh Y, Chuang L, Jiang Y, Chang T, Yang C, Yang C, et al. Application of deep learning image assessment software VeriSee™ for diabetic retinopathy screening. J Formos Med Assoc 2020 Apr 16 [FREE Full text] [doi: 10.1016/j.jfma.2020.03.024] [Medline: 32307321]
36. Chen L, Guh J, Wu K, Chen Y, Kuo M, Hwang S, et al. Modification of diet in renal disease (MDRD) study and CKD epidemiology collaboration (CKD-EPI) equations for Taiwanese adults. PLoS One 2014;9(6):e99645 [FREE Full text] [doi: 10.1371/journal.pone.0099645] [Medline: 24927124]

Abbreviations

CGMH: Chang Gung Memorial Hospital
CKD: chronic kidney disease

eGFR: estimated glomerular filtration rate
ESRD: end-stage renal disease
HbA_{1c}: hemoglobin A_{1c}
MDRD: Modification of Diet in Renal Disease
CNN: convolutional neural network
ReLU: rectified linear unit
ROC: receiver operating characteristic
AUC: area under the curve
PPV: positive predictive values
SNEC: Singapore National Eye Center

Edited by G Eysenbach; submitted 13.08.20; peer-reviewed by YP Huang, G Lim; comments to author 29.08.20; revised version received 01.09.20; accepted 30.10.20; published 26.11.20.

Please cite as:
Kang EYC, Hsieh YT, Li CH, Huang YJ, Kuo CF, Kang JH, Chen KJ, Lai CC, Wu WC, Hwang YS
Deep Learning–Based Detection of Early Renal Function Impairment Using Retinal Fundus Images: Model Development and Validation
JMIR Med Inform 2020;8(11):e23472
URL: http://medinform.jmir.org/2020/11/e23472/
doi:10.2196/23472
PMID:33139242

Original Paper

A Human-Algorithm Integration System for Hip Fracture Detection on Plain Radiography: System Development and Validation Study

Chi-Tung Cheng[1*], MD; Chih-Chi Chen[2*], MD; Fu-Jen Cheng[3], MD, PhD; Huan-Wu Chen[4], MD; Yi-Siang Su[1], MSc; Chun-Nan Yeh[5], MD; I-Fang Chung[6,7,8], PhD; Chien-Hung Liao[1], MD, FACS, FICS

[1]Department of Trauma and Emergency Surgery, Linkou Chang Gung Memorial Hospital, Chang Gung University, Taoyuan, Taiwan
[2]Department of Physical Medicine and Rehabilitation, Linkou Chang Gung Memorial Hospital, Chang Gung University, Taoyuan, Taiwan
[3]Department of Emergency Medicine, Kaohsiung Chang Gung Memorial Hospital, Chang Gung University, Taoyuan, Taiwan
[4]Department of Medical Imaging & Intervention, Linkou Chang Gung Memorial Hospital, Chang Gung University, Taoyuan, Taiwan
[5]Department of General Surgery, Linkou Chang Gung Memorial Hospital, Chang Gung University, Taoyuan, Taiwan
[6]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan
[7]Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei, Taiwan
[8]Preventive Medicine Research Center, Taipei, Taiwan
[*]these authors contributed equally

Corresponding Author:
Chien-Hung Liao, MD, FACS, FICS
Department of Trauma and Emergency Surgery
Linkou Chang Gung Memorial Hospital
Chang Gung University
Trauma Center
5 Fuxin Street, Kweishan District
Taoyuan, 33328
Taiwan
Phone: 886 975365628
Email: surgymet@gmail.com

Abstract

Background: Hip fracture is the most common type of fracture in elderly individuals. Numerous deep learning (DL) algorithms for plain pelvic radiographs (PXRs) have been applied to improve the accuracy of hip fracture diagnosis. However, their efficacy is still undetermined.

Objective: The objective of this study is to develop and validate a human-algorithm integration (HAI) system to improve the accuracy of hip fracture diagnosis in a real clinical environment.

Methods: The HAI system with hip fracture detection ability was developed using a deep learning algorithm trained on trauma registry data and 3605 PXRs from August 2008 to December 2016. To compare their diagnostic performance before and after HAI system assistance using an independent testing dataset, 34 physicians were recruited. We analyzed the physicians' accuracy, sensitivity, specificity, and agreement with the algorithm; we also performed subgroup analyses according to physician specialty and experience. Furthermore, we applied the HAI system in the emergency departments of different hospitals to validate its value in the real world.

Results: With the support of the algorithm, which achieved 91% accuracy, the diagnostic performance of physicians was significantly improved in the independent testing dataset, as was revealed by the sensitivity (physician alone, median 95%; HAI, median 99%; $P<.001$), specificity (physician alone, median 90%; HAI, median 95%; $P<.001$), accuracy (physician alone, median 90%; HAI, median 96%; $P<.001$), and human-algorithm agreement [physician alone κ, median 0.69 (IQR 0.63-0.74); HAI κ, median 0.80 (IQR 0.76-0.82); $P<.001$. With the help of the HAI system, the primary physicians showed significant improvement in their diagnostic performance to levels comparable to those of consulting physicians, and both the experienced and less-experienced physicians benefited from the HAI system. After the HAI system had been applied in 3 departments for 5 months, 587 images were examined. The sensitivity, specificity, and accuracy of the HAI system for detecting hip fractures were 97%, 95.7%, and 96.08%, respectively.

Conclusions: HAI currently impacts health care, and integrating this technology into emergency departments is feasible. The developed HAI system can enhance physicians' hip fracture diagnostic performance.

(JMIR Med Inform 2020;8(11):e19416) doi:10.2196/19416

KEYWORDS

hip fracture; neural network; computer; artificial intelligence; algorithms; human augmentation; deep learning; diagnosis

Introduction

Deep learning (DL) is a subset of machine learning that uses an advanced form of artificial neural networks; the use of DL has impacted health care [1,2]. Numerous applications of DL in medicine, such as computer-aided diagnosis, have been studied [3-9].

Several studies have shown the possibility of using algorithms trained with a large amount of data to aid in appropriate triage, accurately predicting outcomes, improving diagnoses and referrals in clinical situations, and even shortening the waiting time for reports [10-15]. An increasing amount of supporting evidence shows that the use of computer vision with deep neural networks—a rapidly advancing technology ideally suited to solving image-based problems—achieves excellent performance, comparable to that of experts [12-18].

An increasing number of studies have reported the influence of DL in health care, from its use in pathological evaluation to radiographic image assessment [8,9,17,18]. These reports help us understand DL algorithm behavior and how to apply algorithms to reduce medical costs, facilitate further preventive practices, and increase the quality of health care [3].

Hip fractures are among the leading fracture types in elderly individuals worldwide and are the cause of yearly increases in medical costs [19-22]. At the first hospital evaluation, 4-14% of patients' diagnoses are missed [23-25]. Pelvic radiographs (PXRs) are the first-line imaging modality; however, there is a risk of low sensitivity and missed diagnoses when only the image is read. The efficacy and efficiency of several algorithms for skeletal radiology, including hip fracture recognition, have been proven [18,26-28]. However, current state-of-the-art applications of DL to plain film reading by first-line physicians have not been integrated into practice.

Most medical image studies have compared a trained DL algorithm with the performance of human specialists with the goal of developing an algorithm that can outperform specialists [14,26,27,29]. However, there is a lack of studies assessing the combined performance of physician judgment and algorithm prediction, which is a possible situation in clinical scenarios. In this study, we developed a human-algorithm integration (HAI) system to detect hip fractures. Furthermore, we incorporated the HAI system into clinical workflows to improve diagnostic efficiency and accuracy.

Methods

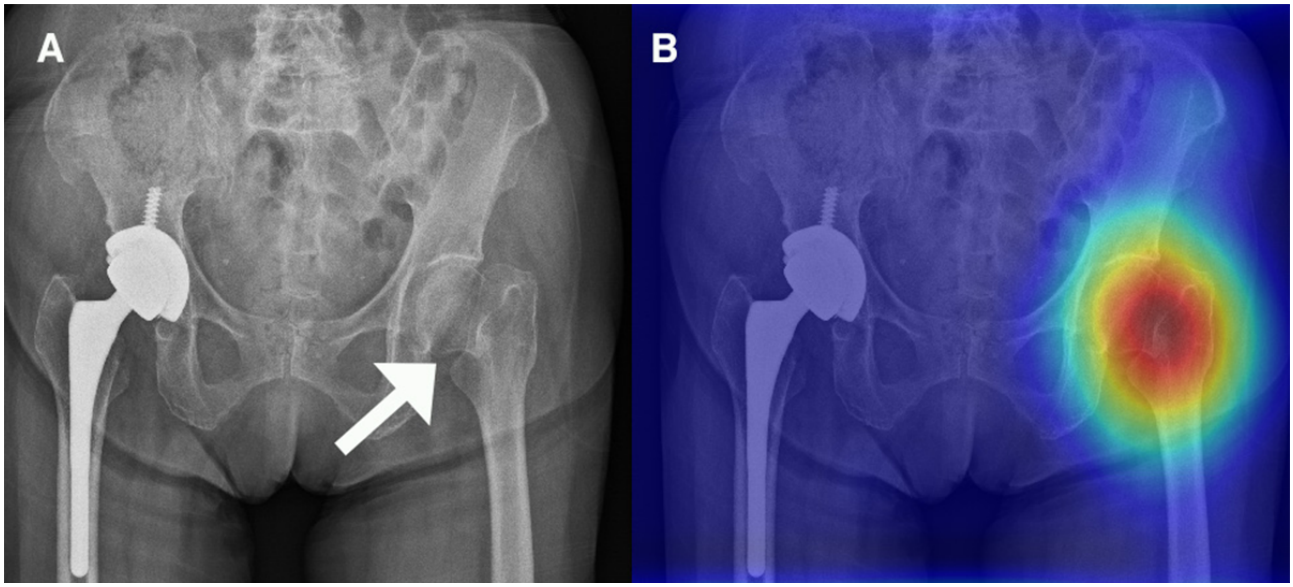
Materials

We utilized data from the Chang Gung Trauma Registry Programme (CGTRP) from Chang Gung Memorial Hospital (CGMH), Linkou, Taiwan. Demographic data, medical records, medical imaging, and associated medical information were recorded prospectively in a computerized database. We extracted the data and images of all trauma patients treated between August 2008 and December 2017 at CGMH, which is a level 1 trauma center. The Internal Review Board of CGMH approved this study. Details of the dataset and image collection process are described in Multimedia Appendix 1.

Development of the Hip Fracture Detection Algorithm and Algorithm Validation

The development of the hip fracture detection algorithm is based on a previous work [18] and described in detail in Multimedia Appendix 1. In summary, we obtained 3605 PXRs, which included 1975 films with hip fractures and 1630 films without hip fractures, from CGMH in Linkou, Taiwan, between August 2008 and December 2016. The diagnostic standard was based on all the available clinical information, including clinical diagnosis, imaging reports, advanced imaging reports, and operative findings. We randomly separated the development dataset into training (2163/3605, 60%), validation (721/ 3605, 20%), and testing (721/3605, 20%) sets for the initial evaluation of the performance of each neural network in hip fracture classification. We assessed VGG16, ResNet-152, Inception-v3, Inception-ResNet-v2, and DenseNet-121 with binary classification with randomly initialized weights. The DenseNet-121 [30] model showed balanced performance regarding the training, validation, and testing sets. Therefore, DenseNet-121 was selected for the classification structure of the deep convolutional neural network (DCNN). We also created heatmaps with gradient-weighted class activation mapping (Grad-CAM) [31] for fracture site detection. We applied the Adam optimizer with an initial learning rate of 10^{-3}. The batch size was 8, and the DCNN was trained for 60 epochs without early stopping. This algorithm was able to evaluate the PXRs and generate a probability of hip fracture and a heatmap overlay for the original image to highlight the possible fracture area, generating an algorithm-assisted reference image for clinician use (Figure 1).

Figure 1. (A) Pelvic x-ray (PXR) with hip fracture. (B) PXR with left hip fracture with a human-algorithm integration–enhanced reference image.



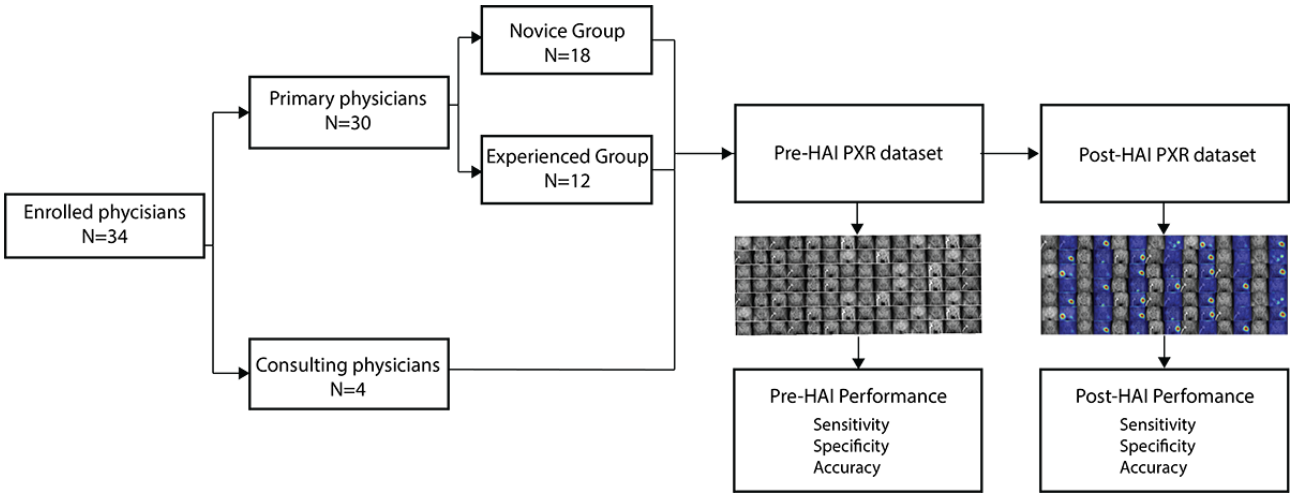
An independent dataset of 100 PXRs (50 hip fracture films and 50 films without hip fractures) from 2017 was collected to evaluate the performance of the algorithm. We set the probability threshold to 0.5, and a cut-off value was also applied in this study. This dataset was used to test the performance of the HAI system.

Study Population, Physician, and the HAI System Performance Test

We enrolled certified medical doctors with different subspecialties and levels of experience and then assessed their performance in an image reading task to validate the HAI system's performance. Subgroup analyses were also performed according to the physicians' experience levels and specialties. Physicians who care for patients in the trauma bay were considered primary physicians, and those who treat patients after consultation (orthopedic surgeons and radiologists) were considered consulting physicians. Based on experience, physicians who had practiced for more than 3 years composed the experienced group; the other physicians composed the novice group.

Before the examination, we introduced the physicians to the study design and provided the image collection details. Figure 2 shows the validation flow of the HAI performance test.

Figure 2. The validation flow of the human-algorithm integration (HAI) system performance test. PXR: plain pelvic radiograph.

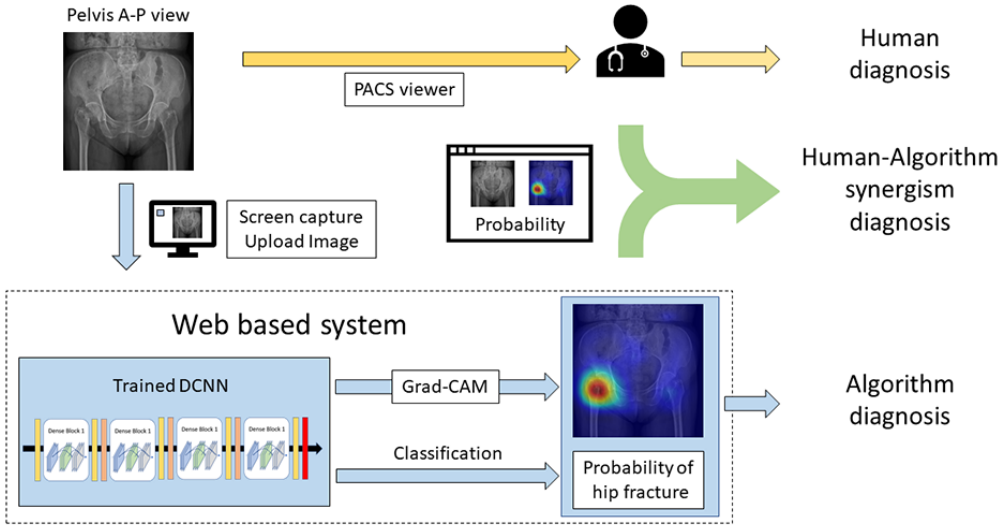


The physicians examined the dataset of 100 PXRs at their original resolution (50 images with fractures and 50 images without fractures) from the validation set from 2017. The physicians were able to zoom in on the images. Upon reviewing the dataset, the physicians were asked to diagnose the presence of a hip fracture. The sensitivity, specificity, and accuracy values obtained composed the physician-alone performance values. Then, the physicians assessed another randomly ordered set of 100 PXRs from the same dataset but with the algorithm-produced reference images, and they were asked to diagnose the presence of a hip fracture. We examined the sensitivity, specificity, and accuracy values of the physicians' HAI performance. We compared the differences in the sensitivity, specificity, and accuracy values based on the algorithm only, the physician's readings only, and the HAI combination. The agreement between the physicians and the algorithm was also calculated on the physician alone and the HAI data.

Real-World Data Study

After validating the feasibility and efficacy of the HAI system, we incorporated the HAI system for physician use in trauma bays and emergency departments at 3 trauma centers in Taiwan: Taipei CGMH, Linkou CGMH, and Kaohsiung CGMH. The physicians could initiate the inference platform of the HAI system while reviewing an image in the picture archiving and communication system (PACS) viewer. The HAI system captured and cropped the PXR from the PACS viewer and transferred it to the central server; the probability of hip fracture was calculated and presented to the clinical physicians along with the original PXR and the reference image (Figure 3).

Figure 3. The flow of clinical integration of the human-algorithm integration (HAI) system into the emergency department and real-world data validation. DCNN: deep convolutional neural network; Grad-CAM: gradient-weighted class activation mapping; PACS: picture archiving and communication system.



All of the images received feedback from the clinical physician to validate whether the diagnosis from the HAI system was correct, and the data were recorded on the server. From the physician's feedback and the clinical diagnosis, we obtained the final report of the accuracy, sensitivity, and specificity of the HAI system. The gold-standard hip fracture diagnosis was the final diagnosis based on all the available clinical information.

Statistical Analysis and Software

The DCNN was built and applied on a machine equipped with the Ubuntu 14.04 operation system (Canonical) with TensorFlow 1.5.1 (Google Brain), Keras 2.1.4, and Keras-vis 0.4.1. Statistical analysis and plots were performed in R 3.4.4 (Microsoft) with the ggplot2 (version 2.0.0; Hadley Wickham) and irr (version 0.84.1; Matthias Gamer et al) packages. Continuous variables were compared using Mann-Whitney U tests and Kruskal-Wallis tests, and categorical variables were evaluated with chi-squared tests. We evaluated the physician-alone and the HAI performance using the sensitivity, specificity, false-negative rate, false-positive rate, and F1 scores; 95% confidence intervals (CIs) were also calculated. Nonnormally distributed data are expressed as medians and interquartile ranges (IQRs). Agreement between the physician and the algorithm was calculated with Cohen kappa. The physician-alone performance and the HAI performance were compared using Wilcoxon signed-rank tests. Receiver operating characteristic (ROC) curves and the areas under the ROC curves (AUCs) were used to evaluate the performance of the model.

Results

DL Algorithm Performance

After applying the hip model to the testing dataset (n=100, normal=50, fractures=50), the sensitivity, specificity, accuracy, and false-negative rate of the model were 98% (95% CI 89%-100%), 84% (95% CI 71%-93%), 91% (n=100; 95% CI 84%-96%), and 2% (95% CI 0.3%-17%), respectively.

Physician and HAI Performance

In total, 34 physicians with a median practice time of 4 (IQR 3.0-5.0) years, including 4 consulting physicians (2 radiologists and 2 orthopedic surgeons) and 30 primary physicians [21 surgeons, 6 emergency physicians, and 3 postgraduate-year (PGY) doctors], completed the examination, as shown in Table 1.

Table 1. Demographic data of the physician participants (n=34).

Physician characteristics	Values
Age in years, median (IQR)	29.00 (27.00-32.00)
Years of practice, median (IQR)	3.00 (2.00-5.75)
Gender, n (%)	
Male	27 (79)
Female	7 (21)
Physician subspecialties, n (%)	
General surgeon	21 (62)
Emergency physician	6 (18)
Postgraduate-year doctor	3 (9)
Radiologist	2 (6)
Orthopedic surgeon	2 (6)

Table 2 shows that the median sensitivity of the primary physicians in the physician-alone testing was 95% (IQR 90%-100%), the median specificity was 90% (IQR 82%-94%), and the median accuracy was 90% (IQR 88%-94%). The median kappa between the physicians and the algorithm was 0.69 (95% CI 0.63-0.74).

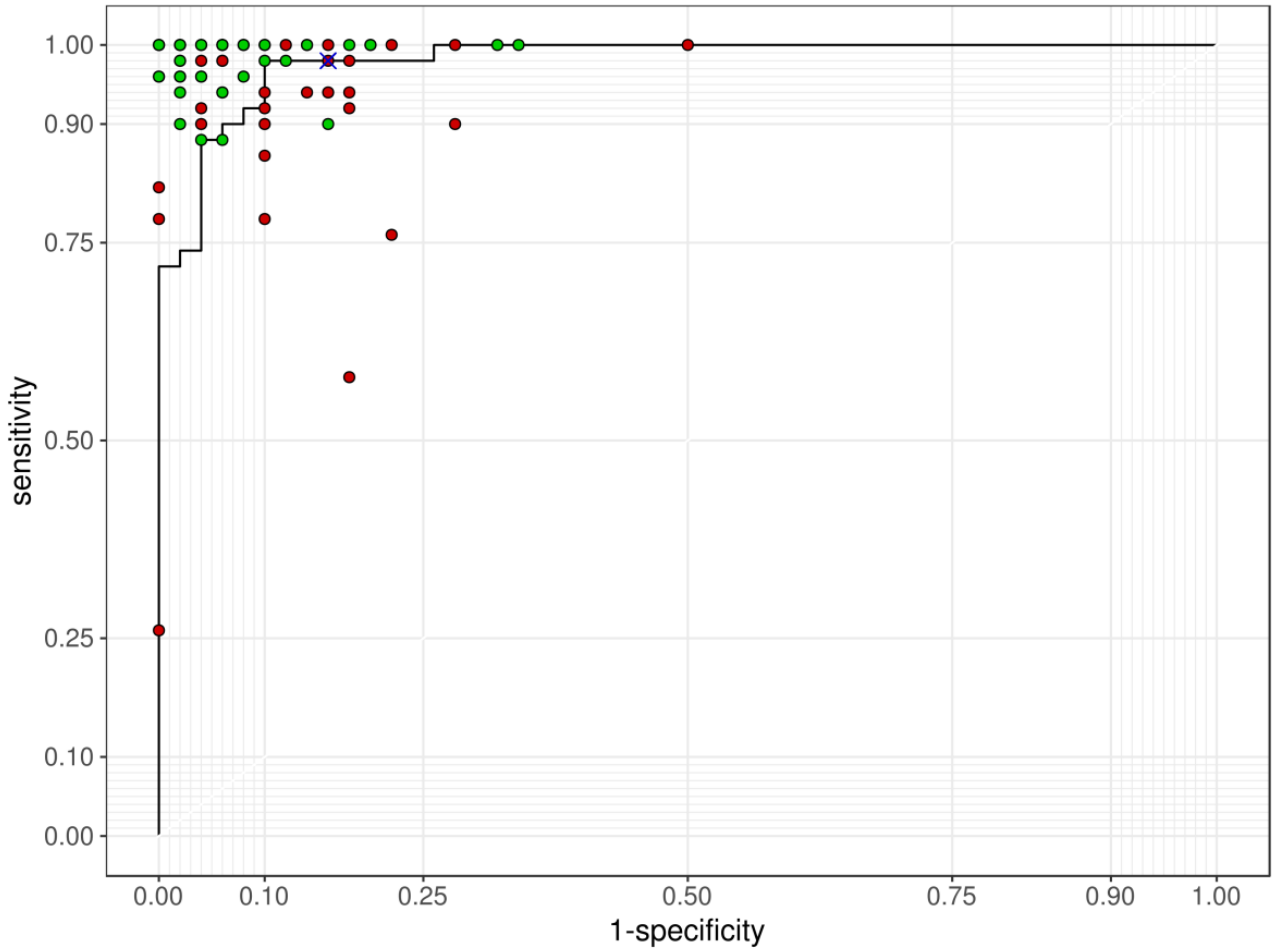
Table 2. The physician-alone performance and human-algorithm integration (HIA) performance of the physician participants (n=34); the Wilcoxon signed-rank test was used to compare the physician-alone and the HAI performance.

Measures	Physician-alone performance	HAI performance	*P* value[a]
Sensitivity, median (IQR)	0.95 (0.90-1.00)	0.99 (0.96-1.00)	<.001
Specificity, median (IQR)	0.90 (0.82-0.94)	0.95 (0.90-0.98)	<.001
Accuracy, median (IQR)	0.90 (0.88-0.94)	0.96 (0.93-0.98)	<.001
Human-algorithm agreement, κ, median (IQR)	0.69 (0.63-0.74)	0.80 (0.76-0.82)	<.001

[a]All *P* values are statistically significant.

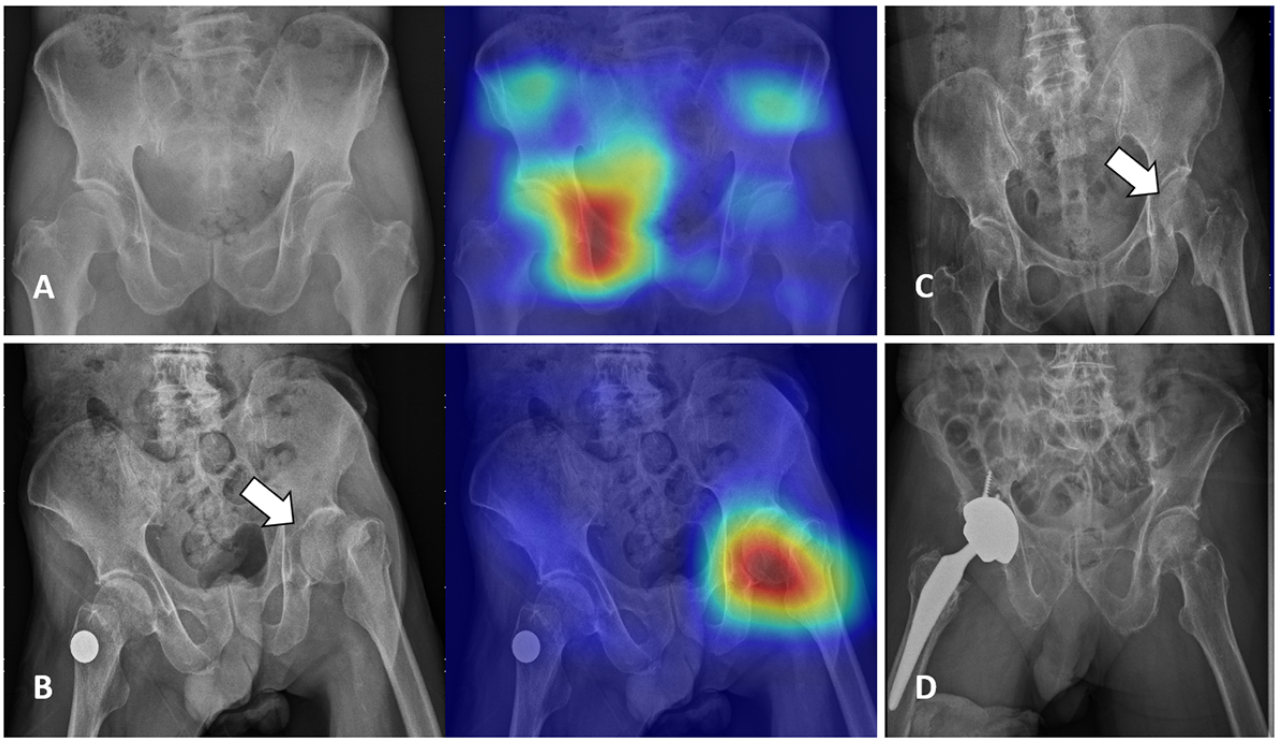
After the HAI system was applied, the median sensitivity of the HAI system was 99% (IQR 96%-100%), the median specificity was 95% (IQR 90%-98%), and the median accuracy was 96% (IQR 93%-98%). The median kappa between the physicians and the algorithm was 0.80 (IQR 0.76-0.82). All of the above factors were significantly improved after HAI system implementation (Table 2). Most of the physicians' performances improved after the algorithm-assisted test, as shown in Figure 4.

Figure 4. The receiver operating characteristic curve of the algorithm performance on test images. Green spots: participants' performance; red spots: participants' performance with human-algorithm integration (HAI) assistance; cross mark: the cut-off performance of the algorithm presented to the physician.



The agreement between the physicians and the algorithm also increased but was still not entirely consistent. Among all the HAI system results, the algorithm had a false-positive rate of 8% per questionnaire, compared with a false-positive rate of only 0.91% per questionnaire for the physicians plus the algorithm. On the other hand, 3.76% of the fractures per questionnaire that were not identified on the physician-alone test were correctly identified after the algorithm information was provided, as shown in Figure 5.

Figure 5. Examples of inconsistencies between the participants and the algorithm. (A) A pelvic x-ray (PXR) without hip fracture that was overdiagnosed by the algorithm. No participant overdiagnosed in the physician-alone test, and only 1 (2.9%) participant overdiagnosed in the human-algorithm integration (HAI) test. (B) A PXR with left hip fracture. In the physician-alone test, 12 (35.3%) participants missed this fracture. In the HAI test, only 1 (2.9%) participant missed this fracture. (C) A PXR with left hip fracture that was missed by the algorithm. In the physician-alone test, 4 (11.8%) participants missed this fracture. In the HAI test, 3 (8.8%) participants missed this fracture. (D) A PXR without hip fracture. In the physician-alone test, 18 (52.9%) participants overdiagnosed this image. In the HAI test, only 5 (14.7%) participants overdiagnosed this image.



Furthermore, the results were divided according to the physicians' specialties, as shown in Table 3. The consulting physicians achieved a better performance than the primary physicians. Regarding the HAI performance, although there were still significant differences in the overall accuracy between specialties, there were no significant differences in the sensitivity and specificity.

Table 3. The physician-alone performance and human-algorithm integration (HAI) performance of the physicians by specialty (n=34).

Physician characteristics and performance	Primary physicians			Consulting physicians		*P* value
	General surgeons (n=21)	Emergency-department physicians (n=6)	Postgraduate-year physicians (n=3)	Radiologists (n=2)	Orthopedic surgeons (n=2)	
Age in years, median (IQR)	28.00 (27.00-30.00)	33.50 (29.75-37.25)	26.00 (25.50- 26.50)	39.00 (36.00-42.00)	33.50 (32.75-34.25)	.006[a]
Years of experience, median (IQR)	3.00 (2.00-4.00)	5.50 (2.75-9.00)	1.00 (1.00- 1.00)	6.50 (6.25- 6.75)	12.50 (10.75-14.25)	.003[a]
Physician-alone performance						
Human-algorithm agreement, κ, median (IQR)	0.69 (0.63-0.72)	0.75 (0.67- 0.80)	0.44 (0.32- 0.53)	0.79 (0.78- 0.79)	0.72 (0.71- 0.74)	.027[a]
Accuracy, median (IQR)	0.90 (0.88-0.92)	0.95 (0.91- 0.97)	0.70 (0.66- 0.76)	0.96 (0.96- 0.97)	0.94 (0.93- 0.94)	.013[a]
Sensitivity, median (IQR)	0.94 (0.90-0.98)	1.00 (0.98- 1.00)	0.58 (0.42- 0.74)	0.99 (0.98- 0.99)	1.00 (1.00- 1.00)	.003[a]
Specificity, median (IQR)	0.90 (0.82-0.96)	0.91 (0.83- 0.94)	0.82 (0.77- 0.91)	0.94 (0.94- 0.94)	0.87 (0.85- 0.88)	.855
HAI performance						
Human-algorithm agreement, κ, median (IQR)	0.80 (0.76-0.82)	0.81 (0.78- 0.83)	0.76 (0.74- 0.78)	0.78 (0.76- 0.80)	0.82 (0.82- 0.82)	.496
Accuracy, median (IQR)	0.95 (0.94-0.97)	0.98 (0.97- 0.99)	0.91 (0.89- 0.91)	0.97 (0.96- 0.97)	1.00 (1.00- 1.00)	.011[a]
Sensitivity, median (IQR)	0.98 [0.96, 1.00]	1.00 (1.00- 1.00)	0.90 (0.89- 0.95)	0.97 (0.95- 0.98)	1.00 (1.00- 1.00)	.121
Specificity, median (IQR)	0.94 (0.90-0.98)	0.97 (0.94- 0.98)	0.84 (0.83- 0.89)	0.97 (0.96- 0.97)	1.00 (1.00- 1.00)	.071

[a]*P* value is statistically significant.

To evaluate the influence of the physicians' clinical experience on the use of the HAI system, we divided the primary physicians into novice and experienced groups, and the results are shown in Table 4. The experienced physicians showed a significantly higher sensitivity and slightly lower specificity than the novice physicians. After the integration of the algorithm information, the overall performance increased regardless of clinical experience.

Table 4. A comparison of the primary physician performance with the human-algorithm integration (HAI) performance, divided by physician experience (n=30); the Wilcoxon signed-rank test was used to compare the physician-alone performance and the HAI performance.

Primary physician characteristics and performance	Novice group (n=18)	Experienced group (n=12)	*P* value
Age in years, median (IQR)	27.00 (27.00-28.00)	32.00 (30.75-34.25)	<.001[a]
Years of experience, median (IQR)	2.00 (2.00-3.00)	5.00 (4.00-6.25)	< .001[a]
Performance evaluation			
Human-algorithm agreement, κ, median (IQR)			
Physician alone	0.66 (0.62-0.72)	0.69 (0.64-0.77)	.330
HAI	0.77 (0.71-0.80)	0.82 (0.79-0.82)	.008[a]
Paired test, *P* value	.0001[a]	.001[a]	
Accuracy, median (IQR)			
Physician alone	0.90 (0.82-0.92)	0.90 (0.89-0.96)	.279
HAI	0.94 (0.91-0.97)	0.97 (0.95-0.98)	.020 [a]
Paired test, *P* value	.0023[a]	.0032[a]	
Sensitivity, median (IQR)			
Physician alone	0.91 (0.83-0.95)	0.98 (0.94-1.00)	.017[a]
HAI	0.97 (0.94-1.00)	1.00 (0.97-1.00)	.043[a]
Paired test *P* value	.0028[a]	.0313[a]	
Specificity, median (IQR)			
Physician alone	0.89 (0.84-0.94)	0.86 (0.81-0.94)	.733
HAI	0.94 (0.88-0.96)	0.96 (0.92-0.98)	.215
Paired test, *P* value	.1067	.0049[a]	

[a]*P* value is statistically significant.

Real-World Validation of the HAI System

In total, 632 tests were completed between March 24, 2019, and August 3, 2019. Images were excluded for the following reasons: (1) poor quality, (2) incorrect image input, such as chest plain film or computed tomography (CT), and (3) PXRs of pediatric patients. After excluding images for the above reasons, 587 PXRs qualified for inclusion. Among the 587 PXRs, there were 320 normal PXRs and 267 PXRs that showed hip fractures. The algorithm's diagnostic accuracy was 92.67% (95% CI 90.26%-94.65%), the sensitivity was 91.01% (95% CI 86.92%-94.16%), the specificity was 94.06% (95% CI 90.88%-96.39%), and the false-negative rate was 7.33%. Of the 587 PXRs, the physicians' diagnoses were consistent with the algorithm for 561 images (95.57%) and were inconsistent for 26 images (4.43%). After reference image assistance, the diagnostic accuracy of the HAI system was 97.10% (95% CI 95.40%-98.30%), the sensitivity was 99.25% (95% CI 97.32%-99.91%), and the specificity was 95.31% (95% CI 92.39%-97.35%). Of the 587 images, 2 images could not be diagnosed by the HAI system; these 2 patients required a CT for hip fracture diagnosis. The false-negative rate of the HAI system was 0.65%, as presented in Table 5.

Table 5. Clinical validation of the human-algorithm integration (HAI) system in emergency departments.

Algorithm-only vs. HIA diagnosis	Fracture		Sensitivity, % (95% CI)	Specificity, % (95% CI)	Accuracy, % (95% CI)
	(+)	(–)			
Algorithm-only diagnosis			91.01 (86.92%-94.16%)	94.06 (90.88%-96.39%)	92.67 (90.26%-94.65%)
(+)	243	19			
(–)	24	301			
HAI diagnosis			99.25 (97.32%-99.91%)	95.31 (92.39%-97.35%)	97.10 (95.40%-98.30%)
(+)	265	15			
(–)	2	305			

Discussion

In this study, we demonstrated 2 findings. First, the HAI system, which integrates an algorithm and human intelligence, performed better than the physicians alone and the algorithm alone. Second, we integrated the HAI system into the clinical flow and verified its use in real-world trauma bays. For orthopedic radiology, fracture detection with computed-aid diagnosis is one of the first applications of AI in radiologic imaging [32,33]. In this study, with the assistance of the HAI system, the physicians detected hip fractures with an increased diagnostic accuracy ranging from 2% to 22%. Several studies demonstrate the strong performance of DL algorithms for fracture detection from different anatomic sites, such as the wrist, humeral, foot, and femur. In this study, our algorithm performance was not inferior to these previous results [26-28]. Furthermore, previous studies usually compared the results of an algorithm with those of professional personnel or other algorithms [17,26,28,34,35].

In the current environment, the algorithm does not replace human intelligence, especially in health care; however, a DL algorithm can complement and augment the ability and knowledge of physicians [1,36,37]. Until now, no real-world data from clinical studies have shown that the integration of AI into the clinical environment can aid physicians. Our study provides the first evidence that HAI can assist patients and doctors in the trauma bay, and we have demonstrated the feasibility of an HAI system to increase diagnostic accuracy.

Some issues occur with the use of computer-assisted diagnostic tools [38,39]. First, the algorithm makes decisions based on features that need to be explored, and there are inevitable caveats, even though the predictions may be correct. Another issue is that physicians may overly rely on the algorithm and disregard their own judgment. To resolve these issues, the HAI system offers physicians a heatmap that highlights the probable location of the fracture on the reference PXR, thus helping physicians understand how the algorithm works. The physician needs to review the image and make the final diagnosis, which prevents him or her from over-relying on the HAI system. We designed a method integrating human expertise and computers that fits the clinical context [38-40], and the HAI system can increase the diagnostic accuracy and specificity. After implementing the validation test performed by 34 physicians, we found that the HAI system performed better than the physicians alone and the algorithm alone (accuracy: 90% vs. 86% vs. 90%; false-negative rate: 6% vs. 12% vs. 9%, respectively). Moreover, we found that with the HAI system, novice physicians can increase their diagnostic accuracy to more closely approach that of experienced physicians, and even consulting physicians.

Machine learning methods have a tendency to “overfit” to idiosyncrasies in the training sample, which may yield overly optimistic performance estimates [36,37]. When addressing the challenges of clinical usage, another question arises: Can the DL algorithm handle real-world data in addition to edited information [1]? Limited studies have proven that algorithms can be applied to real-world data [14], but the clinical effects are still being evaluated. In this study, we have proved that HAI helps physicians detect hip fractures. We operated the HAI system in the trauma bays of 3 trauma centers and obtained adequate hip fracture recognition results. In a real-world validation study, the HAI system improved the accuracy of hip fracture diagnosis to 97%, with a false-negative rate of 0.65%. Several reports have shown that the algorithm might help physicians in acute care and could save lives [10-15].

We did not develop an excellent complete AI solution that can address all situations in health care. However, with the support of HAI, we can reduce some preventable costs and functional losses in fragile fracture cases, improve the allocation of resources, reduce the need for unnecessary consultations, and facilitate faster patient disposition [1,3]. HAI has the potential to improve the delivery of efficient and high-quality care in high-volume clinical practice while allowing physicians to focus on more conceptually demanding tasks by offloading their more mundane duties [3,41].

The clinical usage of the proposed HAI system can improve diagnostic accuracy and reduce the unnecessary use of CT. However, there are still some limitations. First, because we defined our system as an HAI system, selection bias might exist, and clinical physicians could always use this tool when they were unsure about the presence of a fracture. Therefore, some of the images may have been excluded. Second, PXRs provide information on not only skeletal fractures but also soft tissue changes. Although our HAI system can detect fracture sites on PXRs, it still lacks information needed to detect other lesions and cannot replace the expertise of radiologists and clinical physicians. Third, the HAI system does not currently integrate clinical information, which differs from the considerations of clinical practice. The integration of clinical data into the HAI system is another challenge [42]. Fourth, the limited number of physicians participating in this evaluation might have resulted in an underpowered study. Fifth, the images of the validation dataset are only from one institute, which might induce selection bias as well. Finally, a preliminary study of 600 testing images was performed. However, this study tested a limited number of cases at 3 different trauma centers, which are further limitations. The future development of a prospective multicenter study should be used to investigate the system’s function in the real world.

In conclusion, the HAI system improves diagnostic accuracy, and the integration of this technology into the clinical flow is feasible. The HAI system can enhance the performance of physicians, especially novice clinicians.

Acknowledgments

The authors thank CMRPG3H0971, CMRPG3J631, and CIRPG3H0021 for supporting the development of the system and NCRPG3J0012 (MOST109-2622-B-182A-001) for supporting user study. We also thank Mr. Ching-Cheng Chou, the coordinator of Franxense.ai engineering design, and members of Miracle Software Systems for their contribution to the human interface portion of this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The dataset and algorithm development of the Human-Algorithm Integration (HAI) system for hip fracture.
[DOCX File , 29 KB - medinform_v8i11e19416_app1.docx]

References

1. Lynch CJ, Liston C. New machine-learning technologies for computer-aided diagnosis. Nat Med 2018 Sep;24(9):1304-1305. [doi: 10.1038/s41591-018-0178-4] [Medline: 30177823]
2. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018 Feb 22;172(5):1122-1131.e9. [doi: 10.1016/j.cell.2018.02.010] [Medline: 29474911]
3. Berlyand Y, Raja AS, Dorner SC, Prabhakar AM, Sonis JD, Gottumukkala RV, et al. How artificial intelligence could transform emergency department operations. Am J Emerg Med 2018 Aug;36(8):1515-1517. [doi: 10.1016/j.ajem.2018.01.017] [Medline: 29321109]
4. Wong TY, Bressler NM. Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening. JAMA 2016 Dec 13;316(22):2366-2367 [FREE Full text] [doi: 10.1001/jama.2016.17563] [Medline: 27898977]
5. Golden JA. Deep Learning Algorithms for Detection of Lymph Node Metastases From Breast Cancer: Helping Artificial Intelligence Be Seen. JAMA 2017 Dec 12;318(22):2184-2186. [doi: 10.1001/jama.2017.14580] [Medline: 29234791]
6. Yongping L, Juan Z, Zhou P, Yongfeng Z, Liu W, Shi Y. Evaluation of the Quadri-Planes Method in Computer-Aided Diagnosis of Breast Lesions by Ultrasonography: Prospective Single-Center Study. JMIR Med Inform 2020 May 05;8(5):e18251 [FREE Full text] [doi: 10.2196/18251] [Medline: 32369039]
7. Gardezi SJS, Elazab A, Lei B, Wang T. Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review. J Med Internet Res 2019 Jul 26;21(7):e14464. [doi: 10.2196/14464]
8. Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. J Med Internet Res 2018 Oct 17;20(10):e11936 [FREE Full text] [doi: 10.2196/11936] [Medline: 30333097]
9. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018 May;73(5):439-445. [doi: 10.1016/j.crad.2017.11.015] [Medline: 29269036]
10. Kwon J, Jeon K, Kim HM, Kim MJ, Lim S, Kim K, et al. Deep-learning-based out-of-hospital cardiac arrest prognostic system to predict clinical outcomes. Resuscitation 2019 Jun;139:84-91. [doi: 10.1016/j.resuscitation.2019.04.007] [Medline: 30978378]
11. 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]
12. 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). [doi: 10.1186/s13054-019-2351-7]
13. Patel SJ, Chamberlain DB, Chamberlain JM. A Machine Learning Approach to Predicting Need for Hospitalization for Pediatric Asthma Exacerbation at the Time of Emergency Department Triage. Acad Emerg Med 2018 Dec;25(12):1463-1470 [FREE Full text] [doi: 10.1111/acem.13655] [Medline: 30382605]
14. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med 2018 Sep;24(9):1337-1341. [doi: 10.1038/s41591-018-0147-y] [Medline: 30104767]
15. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. Ann Emerg Med 2018 May;71(5):565-574.e2. [doi: 10.1016/j.annemergmed.2017.08.005] [Medline: 28888332]
16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 27;521(7553):436-444. [doi: 10.1038/nature14539]
17. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Dec 02;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]

18. Cheng C, Ho T, Lee T, Chang C, Chou C, Chen C, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019 Oct;29(10):5469-5477 [FREE Full text] [doi: 10.1007/s00330-019-06167-y] [Medline: 30937588]
19. Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int 2006 Dec;17(12):1726-1733. [doi: 10.1007/s00198-006-0172-4] [Medline: 16983459]
20. Leslie WD, O'Donnell S, Jean S, Lagacé C, Walsh P, Bancej C, Osteoporosis Surveillance Expert Working Group. Trends in hip fracture rates in Canada. JAMA 2009 Aug 26;302(8):883-889. [doi: 10.1001/jama.2009.1231] [Medline: 19706862]
21. Bliuc D, Nguyen ND, Milch VE, Nguyen TV, Eisman JA, Center JR. Mortality risk associated with low-trauma osteoporotic fracture and subsequent fracture in men and women. JAMA 2009 Feb 04;301(5):513-521. [doi: 10.1001/jama.2009.50] [Medline: 19190316]
22. Lewiecki EM, Wright NC, Curtis JR, Siris E, Gagel RF, Saag KG, et al. Hip fracture trends in the United States, 2002 to 2015. Osteoporos Int 2018 Mar;29(3):717-722. [doi: 10.1007/s00198-017-4345-0] [Medline: 29282482]
23. Hakkarinen DK, Banh KV, Hendey GW. Magnetic resonance imaging identifies occult hip fractures missed by 64-slice computed tomography. J Emerg Med 2012 Aug;43(2):303-307. [doi: 10.1016/j.jemermed.2012.01.037] [Medline: 22459594]
24. Rehman H, Clement RGE, Perks F, White TO. Imaging of occult hip fractures: CT or MRI? Injury 2016 Jun;47(6):1297-1301. [doi: 10.1016/j.injury.2016.02.020] [Medline: 26993257]
25. Vidán MT, Sánchez E, Gracia Y, Marañón E, Vaquero J, Serra JA. Causes and effects of surgical delay in patients with hip fracture: a cohort study. Ann Intern Med 2011 Aug 16;155(4):226-233. [doi: 10.7326/0003-4819-155-4-201108160-00006] [Medline: 21844548]
26. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol 2019 Feb;48(2):239-244. [doi: 10.1007/s00256-018-3016-3] [Medline: 29955910]
27. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthopaedica 2017 Jul 06;88(6):581-586. [doi: 10.1080/17453674.2017.1344459]
28. Gale W, Oakden-Rayner L. Detecting hip fractures with radiologist-level performance using deep neural networks. arxiv.org. arXiv GCAP, 2017 2017.
29. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018 Dec;24(9):1342-1350. [doi: 10.1038/s41591-018-0107-6] [Medline: 30104768]
30. Huang G, Liu Z, Van DML. Densely Connected Convolutional Networks. CVPR 2017;1(2):3. [doi: 10.1109/cvpr.2017.243]
31. Selvaraju R, 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 IEEE international conference on computer vision 2017:626. [doi: 10.1109/iccv.2017.74]
32. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018 Nov 06;115(45):11591-11596 [FREE Full text] [doi: 10.1073/pnas.1806905115] [Medline: 30348771]
33. Kalmet PHS, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, et al. Deep learning in fracture detection: a narrative review. Acta Orthop 2020 Apr;91(2):215-220 [FREE Full text] [doi: 10.1080/17453674.2019.1711323] [Medline: 31928116]
34. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, the CAMELYON16 Consortium, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 2017 Dec 12;318(22):2199-2210 [FREE Full text] [doi: 10.1001/jama.2017.14585] [Medline: 29234806]
35. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 2017 Dec 12;318(22):2211-2223 [FREE Full text] [doi: 10.1001/jama.2017.18152] [Medline: 29234807]
36. Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. Eur J Intern Med 2018 Feb;48:e13-e14. [doi: 10.1016/j.ejim.2017.06.017] [Medline: 28651747]
37. Liew C. The future of radiology augmented with Artificial Intelligence: A strategy for success. Eur J Radiol 2018 May;102:152-156. [doi: 10.1016/j.ejrad.2018.03.019] [Medline: 29685530]
38. van Hartskamp M, Consoli S, Verhaegh W, Petkovic M, van de Stolpe A. Artificial Intelligence in Clinical Health Care Applications: Viewpoint. Interact J Med Res 2019 Apr 05;8(2):e12100 [FREE Full text] [doi: 10.2196/12100] [Medline: 30950806]
39. Bezemer T, de Groot MC, Blasse E, Ten Berg MJ, Kappen TH, Bredenoord AL, et al. A Human(e) Factor in Clinical Decision Support Systems. J Med Internet Res 2019 Mar 19;21(3):e11732 [FREE Full text] [doi: 10.2196/11732] [Medline: 30888324]
40. Terp S, Seabury SA, Arora S, Eads A, Lam CN, Menchine M. Enforcement of the Emergency Medical Treatment and Labor Act, 2005 to 2014. Ann Emerg Med 2017 Feb;69(2):155-162.e1 [FREE Full text] [doi: 10.1016/j.annemergmed.2016.05.021] [Medline: 27496388]

41. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skeletal Radiol 2019 Aug 4;49(2):183-197. [doi: 10.1007/s00256-019-03284-z]
42. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med 2019;2:31 [FREE Full text] [doi: 10.1038/s41746-019-0105-1] [Medline: 31304378]

Abbreviations

AUC: area under the curve
CT: computed tomography
DCNN: deep convolutional neural network
DL: deep learning
Grad-CAM: gradient-weighted class activation mapping
HAI: human-algorithm integration
PXR: pelvic radiograph
ROC: receiver operating characteristic

Edited by C Lovis; submitted 17.04.20; peer-reviewed by YY Liu, R Kaczmarczyk, E Frontoni, W Sun; comments to author 08.05.20; revised version received 23.05.20; accepted 03.11.20; published 27.11.20.

Please cite as:
Cheng CT, Chen CC, Cheng FJ, Chen HW, Su YS, Yeh CN, Chung IF, Liao CH
A Human-Algorithm Integration System for Hip Fracture Detection on Plain Radiography: System Development and Validation Study
JMIR Med Inform 2020;8(11):e19416
URL: http://medinform.jmir.org/2020/11/e19416/
doi:10.2196/19416
PMID:33245279

Original Paper

Predicting Unplanned Readmissions Following a Hip or Knee Arthroplasty: Retrospective Observational Study

Ramin Mohammadi[1], PhD; Sarthak Jain[1], BT; Amir T Namin[1], PhD; Melissa Scholem Heller[1], BA; Ramya Palacholla[2], MD; Sagar Kamarthi[1], PhD; Byron Wallace[1], PhD

[1]Northeastern University, Boston, MA, United States

[2]Tufts University School of Medicine, Boston, MA, United States

Corresponding Author:
Byron Wallace, PhD
Northeastern University
2208
177 Huntington Ave
Boston, MA,
United States
Phone: 1 6173732402
Email: b.wallace@northeastern.edu

Abstract

Background: Total joint replacements are high-volume and high-cost procedures that should be monitored for cost and quality control. Models that can identify patients at high risk of readmission might help reduce costs by suggesting who should be enrolled in preventive care programs. Previous models for risk prediction have relied on structured data of patients rather than clinical notes in electronic health records (EHRs). The former approach requires manual feature extraction by domain experts, which may limit the applicability of these models.

Objective: This study aims to develop and evaluate a machine learning model for predicting the risk of 30-day readmission following knee and hip arthroplasty procedures. The input data for these models come from raw EHRs. We empirically demonstrate that unstructured free-text notes contain a reasonably predictive signal for this task.

Methods: We performed a retrospective analysis of data from 7174 patients at Partners Healthcare collected between 2006 and 2016. These data were split into train, validation, and test sets. These data sets were used to build, validate, and test models to predict unplanned readmission within 30 days of hospital discharge. The proposed models made predictions on the basis of clinical notes, obviating the need for performing manual feature extraction by domain and machine learning experts. The notes that served as model inputs were written by physicians, nurses, pathologists, and others who diagnose and treat patients and may have their own predictions, even if these are not recorded.

Results: The proposed models output readmission risk scores (propensities) for each patient. The best models (as selected on a development set) yielded an area under the receiver operating characteristic curve of 0.846 (95% CI 82.75-87.11) for hip and 0.822 (95% CI 80.94-86.22) for knee surgery, indicating reasonable discriminative ability.

Conclusions: Machine learning models can predict which patients are at a high risk of readmission within 30 days following hip and knee arthroplasty procedures on the basis of notes in EHRs with reasonable discriminative power. Following further validation and empirical demonstration that the models realize predictive performance above that which clinical judgment may provide, such models may be used to build an automated decision support tool to help caretakers identify at-risk patients.

(JMIR Med Inform 2020;8(11):e19761) doi:10.2196/19761

KEYWORDS

deep learning; natural language processing; electronic health records; auto ML; 30-days readmission; hip arthroplasty; knee arthroplasty

Introduction

Approximately 60% of total hip arthroplasties (THAs) and total knee arthroplasties (TKAs) are covered by Medicare nationwide. The Centers for Medicare and Medicaid Services have focused on total joint replacements as a high-volume and high-cost procedure that should be monitored for cost and quality control [1]. Therefore, bundled payment programs have been proposed to decrease the cost of procedures, shorten length of stay, and reduce the number of readmissions and revision surgeries for THAs and TKAs without sacrificing quality of care [2,3]. Accordingly, bundled payment programs penalize service providers for unscheduled or preventable readmissions [4]. In Massachusetts, for example, Medicare penalized 78% of hospitals for unscheduled readmissions between 2015 and 2016 [5]. In this case, the average penalty for hospitals was 0.7% of the Medicare reimbursement [5]. Models that can identify patients at high risk of readmission might help reduce the total costs and may also improve patient outcomes.

The increase in the use and availability of electronic health records (EHRs) has encouraged researchers to develop and evaluate predictive machine learning (ML) models exploiting EHRs. ML models built over EHRs have now been explored for many clinical predictive tasks, including diagnosis, classification, risk stratification, and medical event prediction [6-9]. A survey of this work is available in a study by Shickel et al [10].

Concerning predicting readmission, Shadmi et al [11] developed a model for 30-day readmission using manually crafted features derived from preadmission data. Similarly, Cai et al [12] used logistic regression (LR) to predict readmission and other outcomes for hospitalized patients. Nguyen et al [13] demonstrated that incorporating EHR data from the full hospital stay can improve 30-day readmission prediction, as compared with incorporating EHR data from the day of admission alone. The difference between our work and these previous efforts is that we are specifically concerned with predicting readmissions following *surgery*, rather than in general, which suggests a more focused approach and evaluation.

The idea of using ML to predict the risk of complications in patients following surgery goes back at least a few decades [14]. Recent efforts have demonstrated the general feasibility of predicting target postoperative complications [15,16]. We do not attempt to exhaustively review these efforts. To the best of our knowledge, none of these efforts have taken an exclusively data-driven approach, without the need for manual feature extraction, to predict the risk of any complications leading to readmission following hip or knee arthroplasty. We aim to address this gap in the literature. These predictions can be made passively and automatically with data from EHRs. If shown superior to direct clinical judgments, these predictions might eventually assist prioritization of proactive care and potentially mitigate complications that lead to readmissions.

This is important, partly because of the high volume of surgeries. In 2017, 700,000 knee replacement procedures were performed in the United States, and this number is likely to increase to 3.48 million surgeries by 2030 [17]. Given the rapid increase in the number of arthroplasty procedures, the need for quality and cost control in general and reducing readmissions and revision surgeries is increasingly clear. Readmissions occur for many reasons, but the 3 most common causes for readmission are surgical site infection, ileus or obstruction, and bleeding [4,17,18].

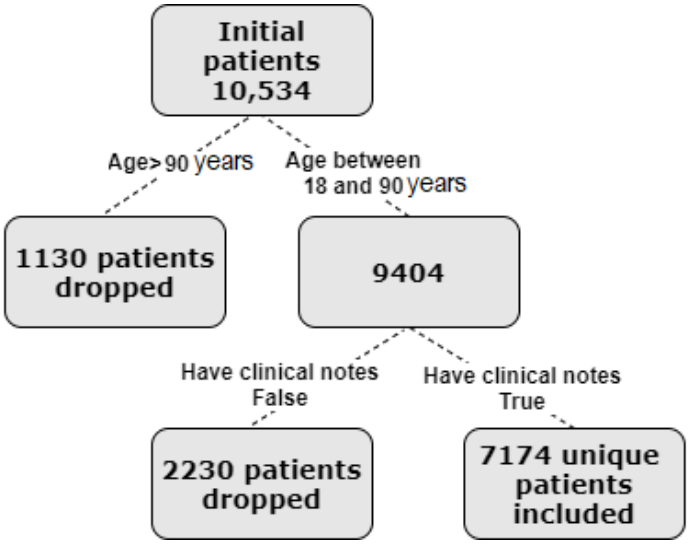
As noted above, there have been previous efforts to predict readmission risk following hip or knee surgery; however, these have relied on structured predictors manually entered by domain experts. This feature extraction process is onerous and precludes automatic and passive monitoring to identify at-risk patients. Our main contribution in this work is the development and evaluation of models for predicting postsurgery readmission directly from EHRs using unstructured clinical notes. In addition, we explored whether neural models induced over clinician notes perform as well or better than simple LR models induced over structured tabular data in the EHRs.

Methods

Data Set

This is a retrospective analysis for which we used EHR data corresponding to 10,534 patients. We received approval from the institutional review board (protocol number 2016P002062 at Partners Healthcare) to conduct this analysis. Subjects were adults aged 18 years or older who were admitted for hip or knee surgery between 2006 and 2016 for either inpatient or outpatient care. These subjects were covered by Medicare, Medicaid, or a private insurance. Our analysis included patients who underwent hip arthroplasty (current procedural terminology [CPT] codes: 27130, 27132, 27134, 27236, 27137, 27138, 27120, and 27125) or knee arthroplasty (CPT codes: 27445, 27446, 27447, 27486, and 27487) during this period. This yielded a data set comprising 7174 patients (Figure 1).

Figure 1. Cohort selection flow chart.



We excluded patients who were aged above 90 years at the time of surgery because of the inherent high risk of complications [19-21], implying that no model is needed for these cases. We also excluded patients for whom no notes were present, which may have induced a sample bias, although we do not have reason to believe this is the case. Figure 1 provides a cohort selection flowchart.

Data Types

Our models exploited (textual) clinical notes to inform predictions. We also considered the use of structured data elements within EHRs for comparison, but we encoded this automatically without domain and ML experts in the loop.

Data Extraction and Encoding

Our primary data set consisted of clinical notes written by clinicians (doctors, nurses, and other health care professionals). These notes described patient demographics, procedures, surgeries, medications, and other medical services rendered to patients. In addition to the free text, notes sometimes contained automatically generated tables (eg, list of laboratory tests). Textbox 1 shows the EHR fields that were considered. In addition to the notes corresponding to these, we often had corresponding structured information. We describe how we preprocess this in the following section.

Textbox 1. Categories of features from electronic health record data used.

Patient level:

- Demographic information
- Health history
- Health information
- Vital information
- Laboratory test results
- Comorbidities
- Medication information
- Radiology
- Procedures
- Surgical
- Pathology
- Diagnosis

Hospital level:

- Admission information

Encounter (visit)

Structured Data Preprocessing

We extracted information pertaining to demographic, diagnostic, encounter, health history, procedures, and medications from patients' records (Textbox 1). Patient encounters are associated with multiple diagnosis codes, including principal, secondary, and other diagnoses. We considered all diagnosis codes when determining whether a readmission was due to surgical complications. We encoded medications and diagnoses as sparse indicator vectors. To process diagnosis International Classification of Diseases (ICD) codes, we mapped ICD-9 codes to ICD-10. We retained only the first 3 ICD-10 characters to reduce sparsity.

To encode variables extracted from the health history table, we concatenated one-hot indicator vectors for all categorical features with numerical values. We encoded laboratory tests using indicator vectors that represent whether a patient received a specific test. For information pertaining to patient health history, we excluded variables that were missing from nearly all (≥99.9%) records (listed in the Multimedia Appendix 1). We also encoded patient medications using indicator vectors. We extracted admission-related information from encounter records (eg, visiting information from admission and discharge sources). For continuous variables, we replaced missing values with averages taken over all patients or encounters as appropriate. This extraction and preprocessing yields, for each patient *z*, *T* encounter records that encode structured elements ordered by the encounter data, where .

Clinical Text Processing

Patients are associated with a list of free-text notes ordered by the encounter date. We tokenize them, then lowercase and stem words, which are then represented via indicator vectors (*V*). All notes are concatenated with a special delineating marker <NOTESEP>, yielding a single note of size where , L_t represents the number of words in note for encounter *t*.

Task Definition

We partitioned the data set at the patient level into train, validation, and test sets with a ratio of 70:15:15 (Table 1). These sets are mutually exclusive with respect to patients (ie, the same patient never appears in more than one set). Demographic statistics for training, validation, and testing sets are reported in the Multimedia Appendices 2-4, respectively. We defined the set of patients who experienced complications following surgery that led to readmission within 30 days using ICD codes. Specifically, we define this as the set of patients who underwent hip or knee surgery and who were subsequently admitted as inpatients within 30 days of their discharge under any of the ICD-9 and ICD-10 complication codes: ICD-9 codes: 996, 996 {03,1-4,57,6,66,67,7,71-73,75-79}, 997, 998 and ICD-10 codes: T84.{0X-7X, 81-86,89,9X}XA.

Table 1. The number of patients in training, validation, and testing data sets.

Data sets	Hip		Knee	
	Male (n=1641), n	Female (n=1658), n	Male (n=1702), n	Female (n=2173), n
Train	1131	1190	1164	1481
Validation	262	238	267	335
Test	248	230	271	357

We labeled patients who met this criterion as having been readmitted due to complications following surgery (*y*=1). We assumed that *all other patients were not readmitted due to complications* (*y*=0). There is an inherent *class imbalance* [22] here; most patients do not experience complications that lead to readmission, that is, there are far fewer positive than zero instances. We report readmission prevalence for hip and knee surgeries in Table 2.

Table 2. Proportion of positive class (30-day readmission because of surgery complications) for hip and knee surgeries.

Subset	Hip	Knee
Train	0.092	0.097
Validation	0.122	0.1
Test	0.115	0.116

Models

We evaluated 2 standard neural models trained on the data set, detailed below. In addition, we implemented a simple LR model to serve as a reference.

Text is encoded into fixed-size representations for downstream modules using an *encoder*. We experimented with a few such encoders: Simple and unstructured count-based bag-of-word (BoW) representations (analogous to the indicator vectors encoding tests and medications) and neural encoders that operate over embeddings of text and learn to represent notes via repeated projection or recurrent modules.

Linear Models (Over Bag of Words)

For our linear model, we used l_1- and l_2-regularized LR over BoW representations of patient notes or the structured data associated with a given patient encounter. We considered 4 different representations of patient notes and structured data associated with a given patient encounter.

BoW variants:

- Binary BoW encodes the existence of a given word in a note as a one-hot vector.
- Count BoW encodes the total number of occurrences of a given word in a note, that is,  .
- Term frequency–inverse document frequency scales word counts inversely to the frequency with which they appear in documents, emphasizing comparatively rare words.
- Finally, we experimented with encoding text via inferred topic distributions using Latent Dirichlet Allocation (LDA) [23]. In this variant, we encoded texts as vectors that encode the proportions of (latent) topics present within them, as estimated via LDA. We report results for LR models that fit text and structured data.

Neural Encoders

Standard neural models first project words to lower-dimensional embeddings (eg, 300 dimensions initialized to pretrained embeddings). These embeddings are then passed through an encoder module before making predictions. We considered the following modules for inducing fixed-length representations of embedded variable-sized textual inputs:

1. Average: Project and then average inputs. Specifically, we first passed embeddings through a linear layer that projects them onto a 256-dimensional space and then applied an element-wise nonlinearity (ReLU).
2. Bidirectional long short-term memory (BiLSTM) network: We ran a single-layer BiLSTM [24] model over the embedded sequence using a hidden layer size of 256 (128 dimensions for each direction).

Recurrent networks (such as BiLSTM) yield variable-sized outputs that must be collapsed into a fixed-length vector. To this end, we adopted a standard max-pooling layer over the outputs of the 256 filters or hidden units. We also explored aggregation via attention mechanisms [25], which allowed models to upweight contextualized representations of specific inputs; accordingly, these have greater influence over the induced fixed-length vector. In the standard attention layer, the model learns to score each encoder hidden state h_t for the input token t according to its relevance for the downstream prediction. Scores are normalized into a distribution α, and a fixed-length vector is induced by taking a weighted sum over the hidden states emitted from the RNN:  . We also explored applying attention to the feedforward (projection) encoder.

In addition, we evaluated hierarchical representation learning over clinician notes [26]. Our data contain reports from different visits. Therefore, we can consider two-level representations: visit level and patient level. An encoder can provide a representation of individual visits, and then these encoded segments can be combined (eg, via a second recurrent neural network) to form a second-level representation of the patient. The latter summarizes all visits. This is referred to as a hierarchical representation. For this, we pass a single BiLSTM to embed each patient's notes separately (using attention), and then we run another BiLSTM over the aggregated patient-level representation of individual notes (associated with its own attention distribution) to yield a fixed-length vector.

Finally, we presented preliminary results using bidirectional encoder representations from transformers (BERT) [27] as another text encoding strategy. Specifically, we used the clinical BERT [28] instantiation of the model that was trained on clinical notes from the MIMIC III data set. BERT is a deep bidirectional model that conditions on both left and right context to provide contextualized representations of words. BERT and similar large pretrained transformer models [29] have achieved good results across many natural language processing data sets and tasks in general; specifically, they have yielded improvements for 30-day readmission tasks on the MIMIC data set [30].

Class Imbalance

Most patients do not experience complications that result in rehospitalization within 30 days. Therefore, the resulting data sets are *imbalanced*, which can be problematic for standard ML models. We experimented with multiple strategies to counteract the class imbalance, including imposing class weights, undersampling the majority class, and oversampling the minority class. Undersampling provided consistent results across data sets and the period of history considered, whereas other strategies proved unstable.

Multitask Learning

The most straightforward approach to predicting 30-day readmission due to complications following hip and knee arthroplasties would be to treat them as an entirely separate class of surgeries and build independent models for each type of surgery. However, intuitively one might expect the information in EHRs to be similar for complications resulting from the respective types of surgery. We can exploit this to improve predictive performance by using multitask learning [31], in which some parameters are shared between models for related tasks.

Performance Metrics

To quantify the performance of the models in predicting 30-day readmission associated with surgical complications, we used the area under the receiver operating characteristic (AUROC) curve and accuracy, sensitivity, specificity, and precision, also known as positive predicted value (PPV), at particular thresholds. These are calculated using true positive (TP), false positive (FP), true negative (TN), and false negative (FN) as follows:

$$\text{Recall (also known as sensitivity)} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FP)$$

$$\text{Precision (also known as PPV)} = TP/(TP+FP)$$

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

In practice, one would need to select an operational threshold with corresponding sensitivity and specificity appropriate for the intended use of the model.

To quantify model performance independent of a particular choice of threshold, we report precision versus recall, and recall versus (1-specificity) and areas underneath the corresponding curves for these constructed by sweeping thresholds (for predicting 1 vs 0) over the predicted probabilities and record

corresponding metrics. The area under these can be taken as a scalar quantifying model performance.

Experimental Setup

Before any experimentation, we separated the data into training, validation, and testing sets. The validation data were used for tuning the models and for selecting the final candidate model. The testing set was used for the final evaluation but was not used in any way during the model development and tuning.

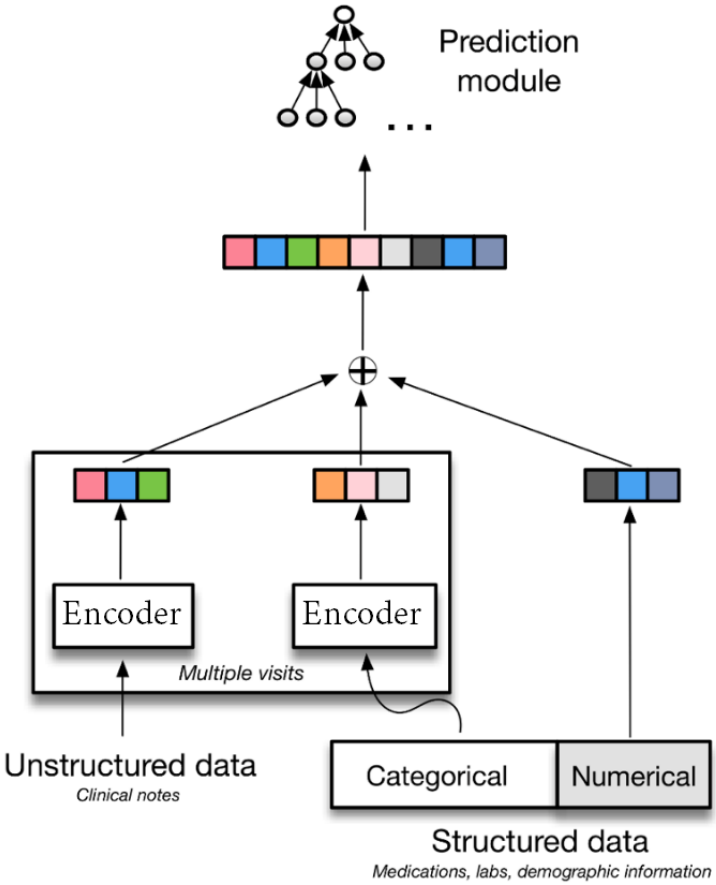
Data Availability

Data supporting this study are not publicly available because of the inherently sensitive nature of the data.

Results

We tuned all hyperparameters on the validation data set. Results achieved under the best models are presented for both hip and knee surgeries as measured on the validation set for (1) text only and (2) structured data only, shown in Figure 2. The results are reported for both the validation and test data sets, where we expect better performance on the former given that we selected hyperparameters based on this. We reported the results for independent models and multitask models over text and structured data separately.

Figure 2. A schematic feature encoding scheme. Structured data, when used, comprises both categorical and numerical elements. We encoded the former using either indicators or an encoder module, whereas we packed the latter into a dense vector of values. Unstructured data (ie, textual notes) are encoded using a sparse (indicator) representation and then optionally run through an encoder module. Colors are stylistics only. The "+" denotes concatenation.



The best independent model for predicting 30-day readmission due to any complications following a hip surgery over the validation data set using text is the feedforward average model with attention mechanism (AUROC=0.894; 95% CI 0.859-0.930); for knee surgeries, the simple feedforward average model performs better (AUROC=0.946; 95% CI 0.929-0.964). Similarly, the best independent model for predicting 30-day readmission due to any complications following hip and knee surgeries using structured data is an LR model with L1 regularization with an AUROC of 0.665 (95% CI 0.589-0.732) and 0.689 (95% CI 0.630-0.749), respectively.

However, the best multitask model for predicting 30-day readmission because of any complications following a hip or knee surgery over text was a feedforward average model with an AUROC of 0.858 (95% CI 0.802-0.915) and 0.937 (95% CI 0.916-0.960), respectively. Similarly, the best multitask model trained over structured data was an LR model with L2 regularization (λ=0.001) with an AUROC of 0.676 (95% CI 0.617-0.738) following hip surgery and an AUROC of 0.664 (95% CI 0.591-0.738) following a knee surgery.

Similarly, the BERT model for predicting 30-day readmission due to any complications following a hip or knee surgery achieved an AUROC of 0.735 (95% CI 0.701-0.785) and 0.820 (95% CI 0.782-0.843), respectively. Therefore, an independent feedforward model over text was selected as the final model to be evaluated for prediction of 30-day unplanned readmission

following knee surgery. Similarly, an independent feedforward model with an attention mechanism developed over text was selected as the model to be evaluated for prediction of 30-day unplanned readmission following hip surgery (Figures 3 and 4).

Figure 3. Precision-recall curve (left) and area under the receiver operating characteristic (AUROC; right) curve for hip validation set. Individual model text (blue), structured (orange), multitask models' text (dashed blue), and structured (dashed orange).

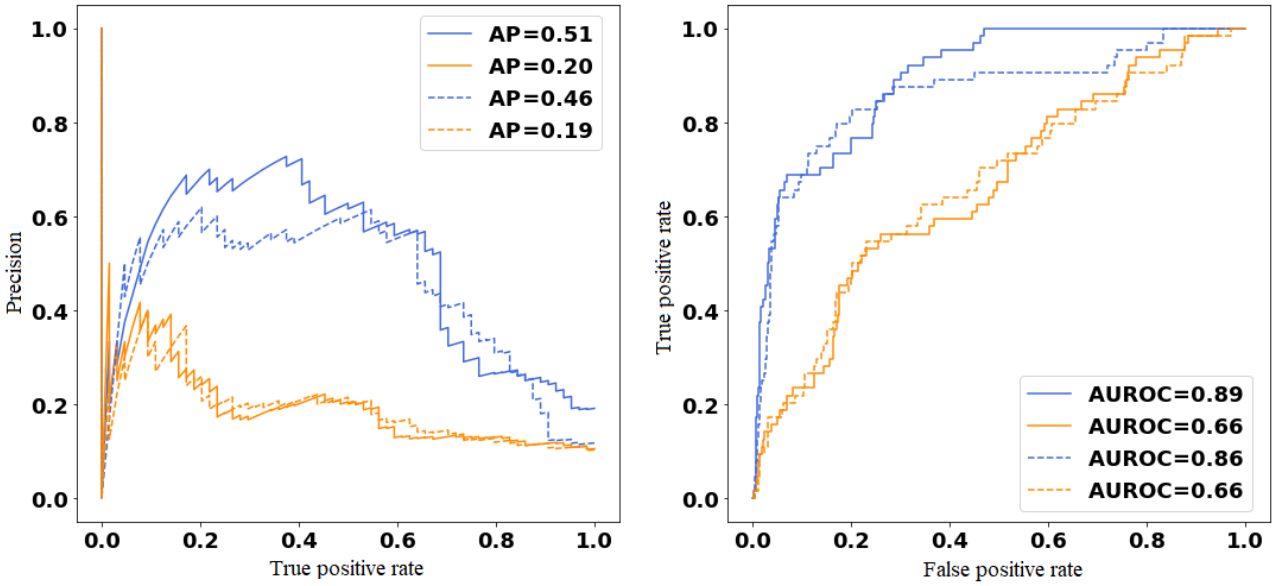
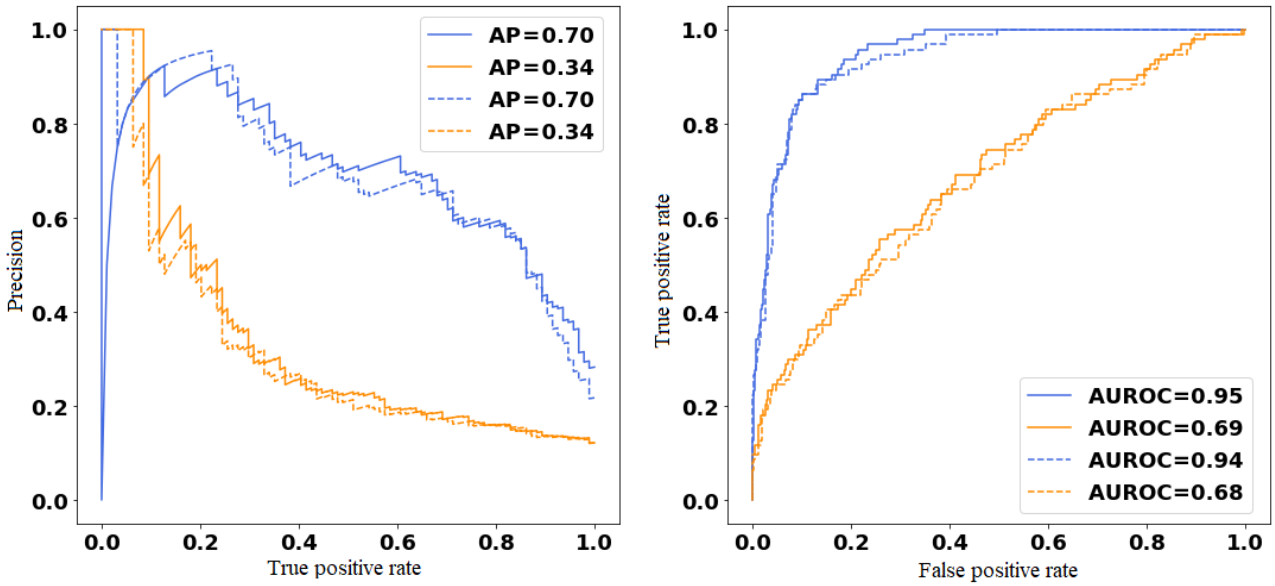


Figure 4. Precision-recall curve (left) and area under the receiver operating characteristic (AUROC; right) curve for knee validation set. Individual model text (blue), structured (orange), multitask models' text (dashed blue), and structured (dashed orange).



We also experimented with a combination of text and structured data (Figure 2). We have not included the results of this experiment in this study because the predictive performance is worse than what we achieved using the text alone. This may seem counterintuitive, but the notes here are relatively rich in information having been manually composed to convey salient information; although these are also noisy at times. It is also entirely possible that alternative feature encodings or model architectures would result in improved model performance with structured data.

We applied the best models (as selected on the validation set) to the test set, realizing an AUROC of 0.846 (95% CI 0.823-0.871) for hip and 0.822 (95% CI 0.809-0.862) for knee surgery.

These AUROCs indicate that the model discriminates between high- and low-risk patients reasonably well. Operationally, such models might conceivably be used to rank patients with respect to their risk of requiring readmission owing to surgical complications and then to provide proactive care (presumably prioritizing limited resources) accordingly. This use would suggest capitalizing on the risk scores and corresponding rankings induced by these directly.

Alternatively, one might seek to establish a binary threshold over model outputs, indicating whether or not action needs to

be taken. The appropriate threshold will depend on the intended use of such a predictive signal, which in turn would depend on the clinical actions at one's disposal and the resources available to take these actions.

Hypothetically, we might entertain 2 settings: first, we prioritize *recall* (ie, *sensitivity*) to identify patients who will need to be readmitted without further intervention at the expense of false-positives and, second, we instead prioritize *precision* (ie, *PPV*) mindful of minimizing false-positives. These 2 settings might correspond, respectively, to a provider who has plentiful resources to provide proactive care (and so false-positives are less of a concern) and a provider who has quite limited resources, which need to be allocated carefully to mitigate false-positive cases.

With this in mind, we selected somewhat arbitrary but illustrative target metrics of 0.95 sensitivity for the former setting and 0.50 precision for the latter. We then selected corresponding thresholds on the validation data and report the results achieved using these on the test data set. Using the first (high-recall) threshold (recall=0.95; precision=0.36 on validation data), the model for readmissions due to complications following knee surgery achieved 0.79 recall and 0.27 precision on test data (classifying *everyone* as positive achieves perfect recall and 0.12 precision). The higher precision threshold (precision=0.50; recall=0.86 on validation data) yields a sensitivity of 0.70 and a precision of 0.40 on test data. For hip surgery, the results are 0.86 recall and 0.22 precision for the high-sensitivity threshold (compared with perfect recall and 0.21 precision) and 0.53 recall and 0.54 precision for the high-precision threshold. We provide results for additional thresholds in Multimedia Appendices 5 and 6.

The clinical utility of such models would, again, depend on how predictions were used in practice.

Related Studies

Previous work has introduced models intended to predict *the risk of readmissions because of complications following colorectal, cardiac, and abdominal surgeries.* For example, Martin et al [32] evaluated predictive factors of hospital readmission rates for 266 patients undergoing abdominal surgical procedures. Wick et al [33] studied the factors associated with readmission using 7 years of data from 10,882 patients who had undergone colorectal surgery. A recent review revealed that the previous predictive models included variables such as patient comorbidities and records of previous hospitalizations [34]. A few other efforts have examined variables associated with severity of illness, laboratory tests, clinical notes from the EMR, and overall health status [33].

The American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) has developed a web-based surgery risk prediction tool that uses structured patient data and LR models to predict risks of complications due to surgery [35]. Edelstein et al [36] evaluated how well ACS-NSQIP can predict 30-day complications following knee and hip replacement surgeries. Mesko et al [37] identified variables predictive of readmission following hip or knee arthroplasty. These approaches rely on a small set of predefined predictors crafted by domain experts that must be manually entered for individuals.

Discussion

Unplanned hospital readmissions impose burdens on the health care system. It is imperative for providers to improve routine follow-up protocols and provide better continuity of care with primary care physicians and other clinicians [38]. Readmission risk prediction models, such as those considered here, might provide insights that could aid decision makers in reducing rehospitalizations and readmissions by identifying patients who might be prioritized to receive proactive care.

Hospital readmissions are a key performance indicator used to measure the quality of care and cost effectiveness of the services provided. In the state of Massachusetts, TKAs and THAs corresponded to a relatively high rate of readmissions from 2010 to 2012 with 3.92% [39]. According to the Nationwide Readmission Database [40] for 224,465 patients participating in the database, the 30-day readmission rate for TKAs is between 3% and 4% depending on Medicare and non-Medicare beneficiaries. A model developed by Urish et al [41] reported that the overall median cost for each 30-day readmission was US $6753 (SD 175), constituting 36% of the overall inpatient cost for 30 days from the index procedures, *which is quite significant.* Clair et al [42] reported the average cost of readmission due to surgical complications after THA and TKA as US $22,775 and US $24,183, for a 90-day readmission with an average readmission time of 31 and 29 days, respectively. The reported costs can be decreased significantly if an appropriate prevention plan is implemented for high-risk patients that are recognized by the adoption of our modeling approach.

We have evaluated several ML algorithms that predict the risk of 30-day readmission following hip and knee arthroplasties by using real-world (unstructured and structured) EHR data obtained from the Partners Healthcare organization. On the basis of the procedure report, the proposed model is able to detect at-risk patients in cases even when there is no sign of complications immediately following the surgery. As evidence for this observation, we reproduce 2 deidentified procedure examples in the Multimedia Appendices 7 and 8.

This study has several limitations, both technical and conceptual. First, we have not evaluated the models' *predictive performance* by comparing predictions with risk of complications of patients as assessed by surgeons or other health care personnel. This may prove to be a strong baseline, but to the best our knowledge, none of the readmission studies used this baseline. However, the fact that risk prediction tools (which rely on manual feature extraction for individual patients) have been studied extensively in this domain suggests a desire for predictive decision aids.

Second, this was a retrospective study using a convenience sample of patient EHR data, which has inherent limitations. Third, although we have demonstrated that ML models can realize reasonably strong overall discriminative performance (in terms of AUROC), translating this into a useful tool in practice would require specifying a threshold that might trigger action. We evaluated a few such hypothetical thresholds but

did not have a clinical basis for these values at the time. However, it is likely that this would depend on the setting in which such models were used.

Fourth, we performed a naïve imputation for missing values, but advanced techniques, including Bayesian [43] and neural system attribution approaches [44], may improve execution. We also excluded variables with a high portion of missing values (≥99%) in patient records; according to domain experts involved in this project, a few of these excluded variables are likely to be clinically relevant. Fifth, we converted the medications and laboratory results into indicator vectors, which may result in information loss, though this was a choice made in consultation with domain experts. Sixth, we used a manually handpicked set of ICD codes to create *labels*, that is, to categorize patients as experiencing complications or not; these ICD codes may be incomplete and may introduce unknown biases in our *positive* samples. Seventh, we excluded patients aged >90 years from our analysis, as we consider such patients to be inherently at high risk.

Finally, the smaller BERT models we used are limited by the size of the document (512 words), whereas most reports here are longer than the limit. In addition, we do not have resources to pretrain BERT on our data set. That said, we tried to follow the clinical BERT methodology to make predictions at the sentence level first and then aggregate the predictions, but this approach did not perform better than the existing neural encoders on our tasks. Although we believe that a more careful application of BERT may result in improvements, it is not a straightforward task, one that needs more research and is not the main goal of the paper.

Conclusions

We presented an ML approach to predict the risk of 30-day readmission following hip or knee arthroplasty using data directly gleaned from EHRs. Previous work on this important problem relied on manually crafted and engineered features, which neither scale nor allow automated surveillance of patients.

We found that our architecture and implementation using the text only (ie, the clinician notes) yielded predictive performance across tasks comparable with approaches using a combination of structured data and text. This suggests that the text contains rich information useful for predicting readmissions. In this case, we also found that adopting a multitask approach (sharing parameters between the models for complications following hip and knee surgeries) did not improve model performance.

We did not aim to identify the specific complication that a patient is comparatively likely to experience. Instead, we offer a patient risk stratification model intended to be used to identify high-risk patients (ie, those most likely to be readmitted) once a clinically meaningful threshold is established. Patients deemed at high risk of readmission because of complications may be scheduled for additional near-term revisits, and in general, be provided with additional proactive care and monitoring. For example, for those identified as high-risk patients, the clinic that is implementing this tool might have a nurse follow-up scheduled for the patient to ensure a continuum of care. This type of risk stratification followed by a nurse intervention in high-risk patients has been shown to produce favorable outcomes, including decreased hospitalizations and cost of care for patients regardless of the complication type [45].

We hope that this initial effort inspires additional work on automatically predicting the risk of readmission because of complications ensuing from hip and knee surgeries because such models have the potential to reduce costs and, more importantly, improve patient outcomes.

Acknowledgments

All the authors have read, edited, and approved the manuscript. The content of the manuscript has not been published or submitted for publication elsewhere.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variables dropped from consideration because of a high proportion of missing values (>99.9%).
[DOCX File , 14 KB - medinform_v8i11e19761_app1.docx]

Multimedia Appendix 2
Training set demographic.
[DOCX File , 17 KB - medinform_v8i11e19761_app2.docx]

Multimedia Appendix 3
Validation set demographic.
[DOCX File , 14 KB - medinform_v8i11e19761_app3.docx]

Multimedia Appendix 4

Testing set demographic.
[DOCX File , 17 KB - medinform_v8i11e19761_app4.docx]

Multimedia Appendix 5
Sample of points on validation set area under the receiver operating characteristic curve for the best model developed on knee surgery.
[DOCX File , 18 KB - medinform_v8i11e19761_app5.docx]

Multimedia Appendix 6
Sample of points on validation set area under the receiver operating characteristic curve for the best model developed on hip surgery.
[DOCX File , 18 KB - medinform_v8i11e19761_app6.docx]

Multimedia Appendix 7
Surgical texts example: surgery with complication.
[DOCX File , 14 KB - medinform_v8i11e19761_app7.docx]

Multimedia Appendix 8
Surgical texts example: surgery without complication.
[DOCX File , 14 KB - medinform_v8i11e19761_app8.docx]

References

1. Namin AT, Jalali MS, Vahdat V, Bedair HS, O'Connor MI, Kamarthi S, et al. Adoption of new medical technologies: the case of customized individually made knee implants. Value Health 2019 Apr;22(4):423-430 [FREE Full text] [doi: 10.1016/j.jval.2019.01.008] [Medline: 30975393]
2. Mears SC, Edwards PK, Barnes CL. How to decrease length of hospital stay after total knee replacement. J Surg Orthop Adv 2016;25(1):2-7. [Medline: 27082881]
3. Hart A, Bergeron SG, Epure L, Huk O, Zukor D, Antoniou J. Comparison of US and Canadian perioperative outcomes and hospital efficiency after total hip and knee arthroplasty. JAMA Surg 2015 Oct;150(10):990-998. [doi: 10.1001/jamasurg.2015.1239] [Medline: 26288005]
4. Desai NR, Ross JS, Kwon JY, Herrin J, Dharmarajan K, Bernheim SM, et al. Association between hospital penalty status under the hospital readmission reduction program and readmission rates for target and nontarget conditions. J Am Med Assoc 2016 Dec 27;316(24):2647-2656 [FREE Full text] [doi: 10.1001/jama.2016.18533] [Medline: 28027367]
5. Unplanned hospital readmissions remain a problem in Mass. Worcester Business Journal. 2016. URL: https://www.wbjournal.com/article/unplanned-hospital-readmissions-remain-a-problem-in-mass [accessed 2019-04-25]
6. Lipton ZC, Kale DC, Elkan C, Wetzell R. Learning to diagnose with lstm recurrent neural networks. arXiv 2015:- epub ahead of print [FREE Full text]
7. Ranganathan R, Perotte A, Elhadad N, Blei D. Deep survival analysis. arXiv 2016:- epub ahead of print [FREE Full text]
8. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
9. Yu C, Fei W, Ping Z. Risk Prediction With Electronic Health Records: a Deep Learning Approach. In: Proceedings of the 2016 International Conference on Data Mining. 2016 Presented at: SIAM'16; December 12-15, 2016; Barcelona, Spain. [doi: 10.1137/1.9781611974348.49]
10. 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. [doi: 10.1109/jbhi.2017.2767063]
11. 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]
12. 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. [doi: 10.1093/jamia/ocv110] [Medline: 26374704]
13. Nguyen OK, Makam AN, Clark C, Zhang S, Xie B, Velasco F, et al. Predicting all-cause readmissions using electronic health record data from the entire hospitalization: model development and comparison. J Hosp Med 2016 Jul;11(7):473-480 [FREE Full text] [doi: 10.1002/jhm.2568] [Medline: 26929062]
14. Lette J, Colletti BW, Cerino M, McNamara D, Eybalin M, Levasseur A, et al. Artificial intelligence versus logistic regression statistical modelling to predict cardiac complications after noncardiac surgery. Clin Cardiol 1994 Nov;17(11):609-614. [doi: 10.1002/clc.4960171109] [Medline: 7834935]

15. FitzHenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. Med Care 2013 Jun;51(6):509-516 [FREE Full text] [doi: 10.1097/MLR.0b013e31828d1210] [Medline: 23673394]
16. Cristina S, Wang MF, Robert J. Data-driven Temporal Prediction of Surgical Site Infection. In: AMIA Annual Symposium Proceedings. 2015 Presented at: AMIA'15; November 12-15, 2015; Chicago, USA.
17. Gregory MM. Patient Education: Total Knee Replacement. UpToDate. 2017. URL: https://www.uptodate.com/contents/total-knee-replacement-beyond-the-basics/print [accessed 2020-10-21]
18. Dixon T, Shaw M, Ebrahim S, Dieppe P. Trends in hip and knee joint replacement: socioeconomic inequalities and projections of need. Ann Rheum Dis 2004 Jul;63(7):825-830. [doi: 10.1136/ard.2003.012724] [Medline: 15194578]
19. SooHoo NF, Lieberman JR, Ko CY, Zingmond DS. Factors predicting complication rates following total knee replacement. J Bone Joint Surg Am 2006 Mar;88(3):480-485. [doi: 10.2106/JBJS.E.00629] [Medline: 16510811]
20. Parvizi J, Mui A, Purtill JJ, Sharkey PF, Hozack WJ, Rothman RH. Total joint arthroplasty: when do fatal or near-fatal complications occur? J Bone Joint Surg Am 2007 Jan;89(1):27-32. [doi: 10.2106/JBJS.E.01443] [Medline: 17200306]
21. Memtsoudis SG, Della Valle AG, Besculides MC, Esposito M, Koulouvaris P, Salvati EA. Risk factors for perioperative mortality after lower extremity arthroplasty: a population-based study of 6,901,324 patient discharges. J Arthroplasty 2010 Jan;25(1):19-26. [doi: 10.1016/j.arth.2008.11.010] [Medline: 19106028]
22. Byron CW, Kevin S, Carla EB. Class Imbalance, Redux. In: 11th International Conference on Data Mining. 2011 Presented at: ICDM'11; December 11-14, 2011; Vancouver,BC, Canada. [doi: 10.1109/icdm.2011.33]
23. David M, Andrew YN, Michael IJ. Latentdirichletallocation. J Mach Learn 2003:1022.
24. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673-2681. [doi: 10.1109/78.650093]
25. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv 2014:- epub ahead of print [FREE Full text]
26. Zichao Y, Diyi Y, Chris D. 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. 2016 Presented at: NAACL HLT'16; June 12-17, 2016; San Diego, California. [doi: 10.18653/v1/n16-1174]
27. Devlin J, Chang M, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv 2018:- epub ahead of print [FREE Full text]
28. Huang K, Jaan A, Ranganathan R. Clinicalbert: modeling clinical notes and predicting hospital readmission. ArXiv 2019:- epub ahead of print [FREE Full text]
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. ArXiv 2017:- epub ahead of print [FREE Full text]
30. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
31. Rich C. Multitask learning. Mach Learn 1997;28(1):75. [doi: 10.1007/978-1-4615-5529-2_5]
32. Martin RC, Brown R, Puffer L, Block S, Callender G, Quillo A, et al. Readmission rates after abdominal surgery: the role of surgeon, primary caregiver, home health, and subacute rehab. Ann Surg 2011 Oct;254(4):591-597. [doi: 10.1097/sla.0b013e3182300a38] [Medline: 22039606]
33. Wick EC, Shore AD, Hirose K, Ibrahim AM, Gearhart SL, Efron J, et al. Readmission rates and cost following colorectal surgery. Dis Colon Rectum 2011 Dec;54(12):1475-1479. [doi: 10.1097/DCR.0b013e31822ff8f0] [Medline: 22067174]
34. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. J Am Med Assoc 2011 Oct 19;306(15):1688-1698 [FREE Full text] [doi: 10.1001/jama.2011.1515] [Medline: 22009101]
35. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg 2013 Nov;217(5):833-42.e1 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2013.07.385] [Medline: 24055383]
36. Edelstein AI, Kwasny MJ, Suleiman LI, Khakhkhar RH, Moore MA, Beal MD, et al. Can the american college of surgeons risk calculator predict 30-day complications after knee and hip arthroplasty? J Arthroplasty 2015 Sep;30(9 Suppl):5-10. [doi: 10.1016/j.arth.2015.01.057] [Medline: 26165953]
37. Mesko NW, Bachmann KR, Kovacevic D, LoGrasso ME, O'Rourke C, Froimson MI. Thirty-day readmission following total hip and knee arthroplasty - a preliminary single institution predictive model. J Arthroplasty 2014 Aug;29(8):1532-1538. [doi: 10.1016/j.arth.2014.02.030] [Medline: 24703364]
38. Brown JR, Sox HC, Goodman DC. Financial incentives to improve quality: skating to the puck or avoiding the penalty box? J Am Med Assoc 2014 Mar 12;311(10):1009-1010 [FREE Full text] [doi: 10.1001/jama.2014.421] [Medline: 24618957]
39. Zawadzki N, Wang Y, Shao H, Liu E, Song C, Schoonmaker M, et al. Readmission due to infection following total hip and total knee procedures: a retrospective study. Medicine (Baltimore) 2017 Sep;96(38):e7961 [FREE Full text] [doi: 10.1097/MD.0000000000007961] [Medline: 28930833]
40. www. NRD Overview. Agency for Health Research and Quality. 2020. URL: https://www.hcup-us.ahrq.gov/nrdoverview.jsp [accessed 2020-11-09]

41. Urish KL, Qin Y, Li BY, Borza T, Sessine M, Kirk P, et al. Predictors and cost of readmission in total knee arthroplasty. J Arthroplasty 2018 Sep;33(9):2759-2763 [FREE Full text] [doi: 10.1016/j.arth.2018.04.008] [Medline: 29753618]
42. Clair AJ, Evangelista PJ, Lajam CM, Slover JD, Bosco JA, Iorio R. Cost analysis of total joint arthroplasty readmissions in a bundled payment care improvement initiative. J Arthroplasty 2016 Sep;31(9):1862-1865. [doi: 10.1016/j.arth.2016.02.029] [Medline: 27105556]
43. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Br Med J 2009 Jun 29;338:b2393 [FREE Full text] [doi: 10.1136/bmj.b2393] [Medline: 19564179]
44. Lipton ZC, Kale DC, Wetzel R. Modeling missing data in clinical time series with rnns. Mach Learn Healthcare 2016:- [FREE Full text]
45. Inouye, Wagner DR, Acampora D, Horwitz RI, Cooney LM, Tinetii ME. A controlled trial of a nursing-centered intervention in hospitalized elderly medical patients: the Yale Geriatric Care Program. J Am Geriatr Soc 1993 Dec;41(12):1353-1360. [doi: 10.1111/j.1532-5415.1993.tb06487.x] [Medline: 8227919]

Abbreviations

ACS-NSQIP: American College of Surgeons National Surgical Quality Improvement Program
AUROC: area under the receiver operating characteristic
BERT: bidirectional encoder representations from transformers
BiLSTM: bidirectional long short-term memory
BoW: bag of words
CPT: current procedural terminology
EHR: electronic health record
FN: false negative
FP: false positive
ICD: International Classification of Diseases
LDA: Latent Dirichlet Allocation
LR: logistic regression
ML: machine learning
PPV: positive predicted value
THA: total hip arthroplasty
TKA: total knee arthroplasty
TN: true negative
TP: true positive

Edited by C Lovis; submitted 30.04.20; peer-reviewed by S Veeranki, J Kim; comments to author 21.06.20; revised version received 08.09.20; accepted 13.09.20; published 27.11.20.

Please cite as:
Mohammadi R, Jain S, Namin AT, Scholem Heller M, Palacholla R, Kamarthi S, Wallace B
Predicting Unplanned Readmissions Following a Hip or Knee Arthroplasty: Retrospective Observational Study
JMIR Med Inform 2020;8(11):e19761
URL: https://medinform.jmir.org/2020/11/e19761
doi:10.2196/19761
PMID:33245283

Original Paper

Machine Learning Electronic Health Record Identification of Patients with Rheumatoid Arthritis: Algorithm Pipeline Development and Validation Study

Tjardo D Maarseveen[1], BSc; Timo Meinderink[2,3], MSc; Marcel J T Reinders[4,5], PhD; Johannes Knitza[2,3], MD; Tom W J Huizinga[1], MD; Arnd Kleyer[2,3], MD; David Simon[2,3], MD; Erik B van den Akker[4,5], PhD; Rachel Knevel[1,6], MD

[1]Department of Rheumatology, Leiden University Medical Center, Leiden, Netherlands

[2]Department of Internal Medicine 3, Friedrich-Alexander University Erlangen - Nuremberg, Erlangen, Germany

[3]Deutsches Zentrum für Immuntherapie, Erlangen-Nuremberg and Universitätsklinikum, Erlangen, Germany

[4]Leiden Computational Biology Centre, Leiden University Medical Center, Leiden, Netherlands

[5]Molecular Epidemiology, Leiden University Medical Center, Leiden, Netherlands

[6]Division of Rheumatology, Inflammation and Immunity, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Rachel Knevel, MD
Department of Rheumatology
Leiden University Medical Center
C1-R k. 41
Albinusdreef 2
Leiden, 2333 ZA
Netherlands
Phone: 31 611307780
Email: R.Knevel@lumc.nl

Abstract

Background: Financial codes are often used to extract diagnoses from electronic health records. This approach is prone to false positives. Alternatively, queries are constructed, but these are highly center and language specific. A tantalizing alternative is the automatic identification of patients by employing machine learning on format-free text entries.

Objective: The aim of this study was to develop an easily implementable workflow that builds a machine learning algorithm capable of accurately identifying patients with rheumatoid arthritis from format-free text fields in electronic health records.

Methods: Two electronic health record data sets were employed: Leiden (n=3000) and Erlangen (n=4771). Using a portion of the Leiden data (n=2000), we compared 6 different machine learning methods and a naïve word-matching algorithm using 10-fold cross-validation. Performances were compared using the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC), and F1 score was used as the primary criterion for selecting the best method to build a classifying algorithm. We selected the optimal threshold of positive predictive value for case identification based on the output of the best method in the training data. This validation workflow was subsequently applied to a portion of the Erlangen data (n=4293). For testing, the best performing methods were applied to remaining data (Leiden n=1000; Erlangen n=478) for an unbiased evaluation.

Results: For the Leiden data set, the word-matching algorithm demonstrated mixed performance (AUROC 0.90; AUPRC 0.33; F1 score 0.55), and 4 methods significantly outperformed word-matching, with support vector machines performing best (AUROC 0.98; AUPRC 0.88; F1 score 0.83). Applying this support vector machine classifier to the test data resulted in a similarly high performance (F1 score 0.81; positive predictive value [PPV] 0.94), and with this method, we could identify 2873 patients with rheumatoid arthritis in less than 7 seconds out of the complete collection of 23,300 patients in the Leiden electronic health record system. For the Erlangen data set, gradient boosting performed best (AUROC 0.94; AUPRC 0.85; F1 score 0.82) in the training set, and applied to the test data, resulted once again in good results (F1 score 0.67; PPV 0.97).

Conclusions: We demonstrate that machine learning methods can extract the records of patients with rheumatoid arthritis from electronic health record data with high precision, allowing research on very large populations for limited costs. Our approach is language and center independent and could be applied to any type of diagnosis. We have developed our pipeline into a universally

applicable and easy-to-implement workflow to equip centers with their own high-performing algorithm. This allows the creation of observational studies of unprecedented size covering different countries for low cost from already available data in electronic health record systems.

(JMIR Med Inform 2020;8(11):e23930) doi:10.2196/23930

KEYWORDS

Supervised machine learning; Electronic Health Records; Natural Language Processing; Support Vector Machine; Gradient Boosting; Rheumatoid Arthritis

Introduction

Electronic health records (EHR) offer an interesting collection of clinical information for observational research, yet a crucial step is an accurate identification of disease cases. This is commonly done by manual chart review or by using standardized billing codes. However, these methods are either labor-intensive or prone to including false positives. Previous studies [1] found that using only standardized billing codes, for example, ≥3 International Classification of Diseases, Ninth Revision (ICD-9) rheumatoid arthritis codes, results in a positive predictive value (PPV) of 56% (95% CI 47%-64%). Using a combination of billing code with a disease-modifying antirheumatic drug code (≥1 ICD-9 rheumatoid arthritis code plus ≥1 disease-modifying antirheumatic drug) results in a PPV of 45% (95% CI 37%-53%). Clinical diagnoses can also be inferred by performing naïve word-matching on format-free text fields. This approach does not take into account the provided context and is thus prone to false positives as well.

Alternatively, query-like algorithms can be used. However, these algorithms require knowledge on the diagnosis of interest, biasing the inclusion of potential study cases. For example, when we want to identify patients with rheumatoid arthritis, we can select people with cyclic citrullinated peptide antibodies that were treated with methotrexate. Those identified likely concern true cases of rheumatoid arthritis but are biased as patients with rheumatoid arthritis do not always receive methotrexate and do not all have cyclic citrullinated peptide–positive tests. On the other hand, selecting only methotrexate would create many false positives as methotrexate is prescribed for many other rheumatic diseases. An additional disadvantage is that rule-based algorithms tend to be center-specific and perform less well in other clinics [2].

Advancements in natural language processing and machine learning have created great potential for processing format-free text data such as those in EHRs [2,3]. A major advantage of machine learning is that it can learn extraction patterns from a set of training examples, relieving the need for extensive domain knowledge. We set out to explore the utility of machine learning methods to identify patients with rheumatoid arthritis from format-free text fields in EHRs. As machine learning methods learn from presented training examples, they can suffer from intercenter variability due to different notation characteristics in EHRs [2].

Therefore, the aim of this study was to develop a broadly applicable workflow that employs machine learning methods to identify patients with rheumatoid arthritis from format-free text fields of EHRs. Additionally, the workflow should be easy to implement and require only the annotation of a subset of the total data set.

Methods

Patients' Data Collection

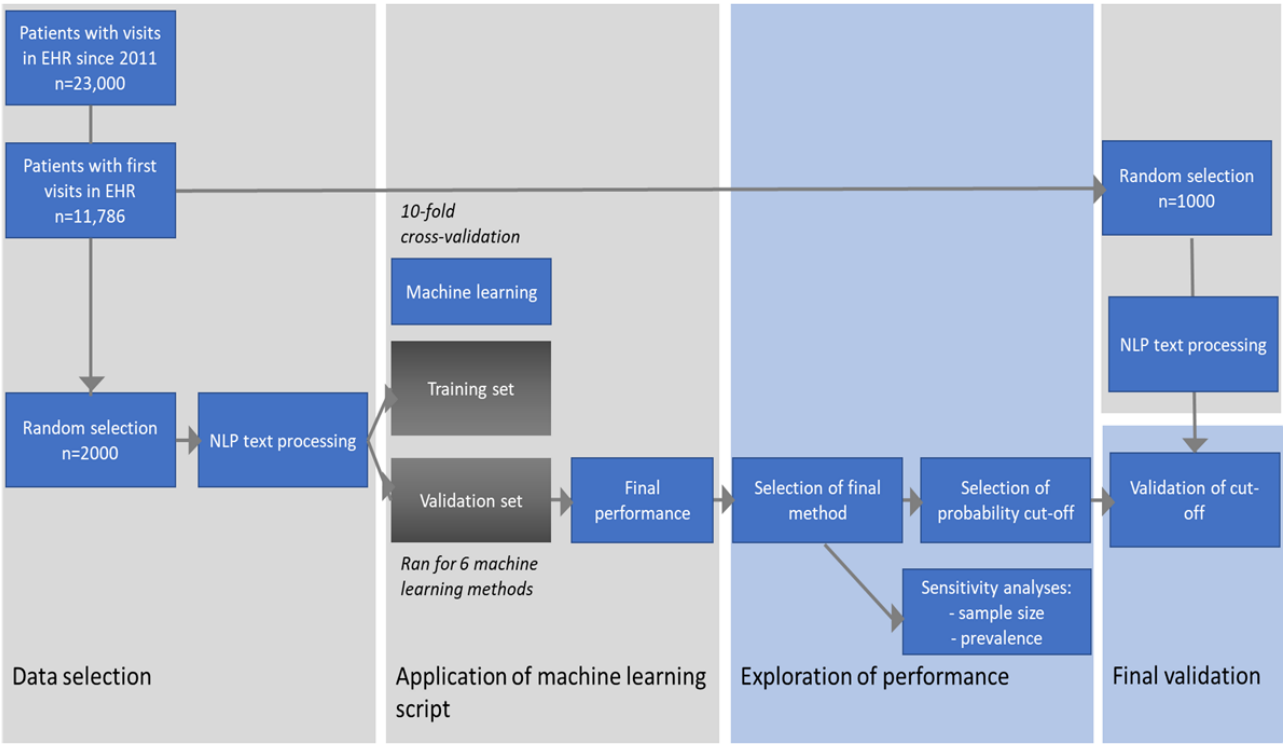
Overview

For this study, we employed 2 data sets: Leiden (the Netherlands) and Erlangen (Germany). See Multimedia Appendix 1 (Table S1) for a convenient overview of the study outline for both centers.

Leiden Data Set

We retrieved EHR data from patients (n=23,300) who visited the rheumatology outpatient clinic of the Leiden University Medical Centre since 2011 (Figure 1). We used the *Conclusion* section of the patient records, which consisted of format-free text fields describing the symptoms and (differential) diagnoses of the patient. From these dossiers, 11,786 patients had a first visit after the initiation of the digital system in 2011 [4]. We randomly selected 3000 patients from these newly referred patients and extracted all of their entries for up to 1 year of follow-up. A clinician manually reviewed all entries and annotated the final diagnosis based on all entries. The data were divided into 2 independent sets with a 66/33 split: Leiden-A (n=2000) for model selection, training, and validation and Leiden-B (n=1000) for independent testing. The study was approved by the local ethics board.

Figure 1. Study outline of the Leiden cohort. EHR: electronic health record; NLP: natural language processing.



Erlangen Data Set

After model selection, training, and validation analyses were performed on the Leiden data, we evaluated the universal applicability of our pipeline by applying it to the EHR data from a second center. We retrieved admission notes from the EHR database of University Hospital Erlangen (Department of Internal Medicine 3 Rheumatology and Immunology, Universitätsklinikum). The *course & assessment* component was used because it featured the patient status descriptions. These data consisted of 4771 patients in total featuring all their entries up to 1 year of follow-up. A health care professional manually reviewed all entries and annotated the final diagnosis based on all entries. The Erlangen data set was divided into 2 independent sets with a 90/10 split: Erlangen-A (n=4293) for model and Erlangen-B (n=478) for testing. The study was approved by the local ethics board.

Training, Model Selection, and Validation (Leiden-A and Erlangen-A)

Preprocessing Format-Free Text

We employed spell check and several natural language processing techniques to preprocess the extracted text with scikit-learn tools provided by Pedregosa et al [5]. The pipeline can be divided into 5 steps: word segmentation, lowercase conversion, stop word removal, word normalization, and vectorization. First, we segmented the text into words, splitting by spaces and special characters. Next, we converted the text to lowercase and removed the irrelevant but highly prevalent stop words. Morphological variation was further reduced by applying lemmatization to normalize words to their base form. The tools provide lemmatization tools for many languages; we used the Dutch and German language tools. Segmented words were then aggregated by grouping neighboring words into sets of 3 (ie, *n*-grams such as *patient*, *verdenking artritis*). Finally, a *term frequency by inverse document frequency* transformation, which builds a clinical vocabulary and weighs words according to their occurrence, was applied to vectorize the text data.

Training and Machine Learning Model Selection

We tested the following machine learning methods: naïve Bayes [6], neural networks [7], random forest [8], support vector machine [9], gradient boosting [10], decision tree [8], and a random classifier, which assigns class labels at random with frequencies equal to those observed in the training set (parameters are shown in Table S2, Multimedia Appendix 1). Default scikit-learn implementations were used to create the machine learning models [11].

Furthermore, we employed a naïve word-matching algorithm that assigns rheumatoid arthritis status to a sample when the text contained rheumatoid arthritis (in German or Dutch) or its abbreviation appeared in the chart. Each classifier gives a score between 0 and 1 that we interpreted as a probability for each sample to be a case.

We randomly split the Leiden-A and Erlangen-A in train and validation sets using a 10-fold cross-validation procedure for model selection [12]. In short, for each sample set, different models were trained and evaluated in equally sized training and validation sets. Classification performances in the validation sets were then averaged over the samples to give robust estimates of each individually evaluated method to annotate unseen EHR records with a rheumatoid arthritis status.

Performance Validation

As each classifier generates a probability score of a rheumatoid arthritis, the performance of a classifier can be tested by applying different cut-offs for case identification. With these

probabilities, we first generated receiver operating characteristic curves, plotting the true positive rate against the false positive rate for all probability scores. Second, we created precision-recall curves, plotting the precision (PPV) against the recall (sensitivity or true positive rate) for all score thresholds. Classification performance was then measured using the area under the receiver operating characteristic curve (AUROC) and the area under the precision curve (AUPRC) [11]. For data sets with low case prevalence (imbalanced data), AUROC can be inaccurate and using AUPRC is preferred [13].

To determine whether the performance of the method significantly differed from that of the word-matching method, we implemented the 5×2 cross-validation procedure described by Dietterich [14]. The 5×2 cross-validation procedure splits the data into 2 equal sized sets each repetition. The differences between the classifiers are then estimated with a two-tailed paired *t* test with a significance level of 0.05. This approach takes into account the problem of dependence between the measurements.

The F1 score served as the primary criterion for picking the final method. The F1 score reflects the trade-off between precision and recall as it is the harmonic mean of the two [15]. The best performing model was compared to the other classifiers with two-tailed paired *t* tests (α=.05) in the 5×2 cross-validation, to evaluate whether the best performing model significantly outperformed the other candidates.

Sensitivity Analyses

We ran 2 sensitivity analyses on the Leiden data. To evaluate the influence of sample size on the performance of a classifier, we employed the classifier on the Leiden-A data set with decreasing sample sizes within the same 10-fold cross-validation setup. To test the effect of disease prevalence on the classifier's performance, we created subsets of the Leiden-A set with different fractions of patients with rheumatoid arthritis, applied the classifier to this data and compared the AUPRC between the subsets.

Final Method Testing of Case Identification (Leiden-B and Erlangen-B)

In the final test phase (using the B data sets), we obtained reliable estimates of the selected model's performance. We applied the trained model for the best performing method from the A data sets directly to the B data sets (Leiden-B, n=1000; Erlangen-B, n=478). To make a final call on rheumatoid arthritis status, one must define a threshold for the probability. The final test characteristics of the model are affected by the chosen probability cut-off. We report the PPV, sensitivity, and F1 score for each B data set at 2 operator points learned from the A data sets: (1) optimized PPV, thus favoring high-certainty cases and (2) optimized sensitivity, thus favoring the inclusive selection of cases.

Implementation and Availability

Machine learning methods, model training, and evaluations were performed with the scikit-learn package (version 0.21.2) in Python (version 3.5) [11]. At all times, default implementations and default settings were used. All scripts including instructions on how to apply the methods are posted online [16].

Results

Data

Leiden-A (n=2000) and Leiden-B (n=1000) annotated data sets had nearly equal percentages of patients with rheumatoid arthritis (Leiden-A: 154/2000, 7.7%; Leiden-B: 84/1000, 8.4%). Erlangen-A (n=4293) and Erlangen-B (n=478) annotated data sets also had nearly equal percentages of patients with rheumatoid arthritis (Erlangen-A: 1071/4293, 24.9%; Erlangen-B: 112/478, 23.4%).

Leiden

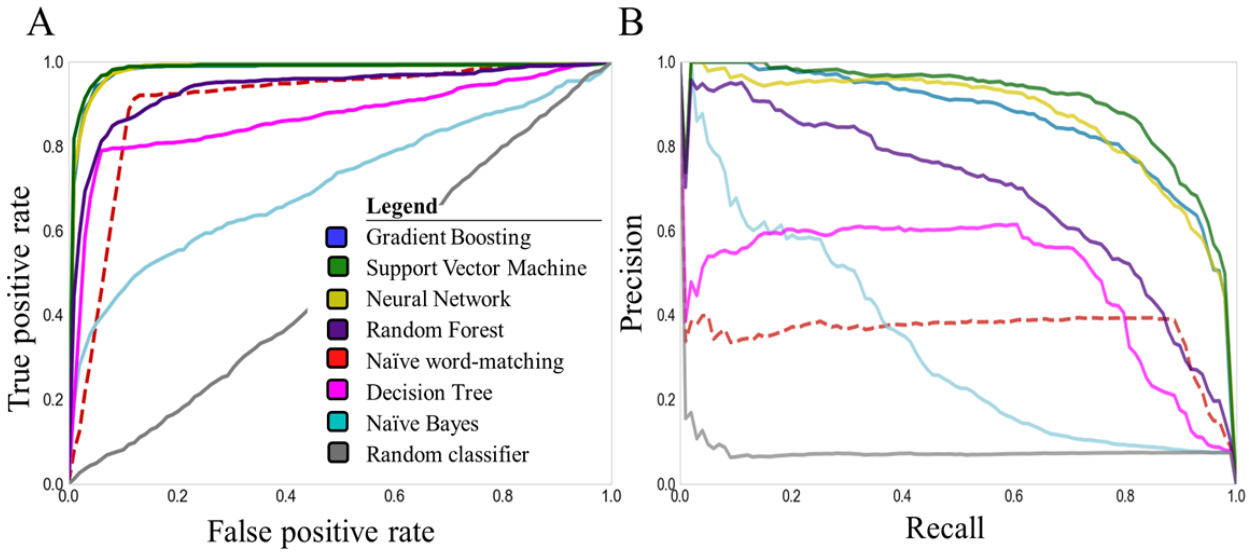
Preprocessing

We found a total of 114,529 words and 8355 unique words in the Leiden-A data after segmentation. With lemmatization and lowercase conversion, the number of unique words was 8141. After removing the most common words with a stop word filter, only 88,524 words and 8078 unique words remained. There were 133,161 unique word combinations (*n*-grams) in the text. The term frequency by inverse document frequency transformation resulted in a sparse matrix of 2000×133,161.

Performance Evaluation of Machine Learning Methods

Naïve word-matching had overall a good performance (AUROC: mean 0.90, SD 0.02), which was significantly better (P<.001) than that of a random classifier (AUROC: mean 0.50, SD 0.01). Although naïve word-matching showed good overall test performance, it had a low AUPRC value (mean 0.36 SD 0.07), indicating that the naïve word-matching would generate many false positives. Four machine learning methods outperformed naïve word-matching (AUROC: naïve Bayes mean 0.71, SD 0.03, P=.003; neural network: mean 0.98, SD 0, P=.005; random forest: mean 0.95, SD 0.01, P=.007; support vector machine: 0.98, SD 0.01, P=.004; gradient boosting: mean 0.98, SD 0.01, P=.003; decision tree: mean 0.86, SD 0.05, P=.06) (Figure 2).

Figure 2. (A) Receiver operating characteristics and (B) precision-recall curves for all machine learning methods (solid lines) and the naïve word-matching method (dotted line) in the training set (Leiden-A).



The support vector machine had the highest performance in comparison to that of word-matching (AUPRC: mean 0.90, SD 0.02; F1 score: mean 0.83 SD 0.02, $P<.001$). However, the 5×2 cross-validation paired *t* tests revealed that the differences for gradient boosting ($P=.61$), neural network ($P=.18$), and random forest ($P=.10$) were not significant (Multimedia Appendix 2).

Sensitivity Analyses

We did not observe any significant loss of precision when lowering the number of training samples from 1000 (original) to 600 patients (Multimedia Appendix 3). Neither the AUROC nor the AUPRC showed a significant difference ($P=.17$ and $P=.11$, respectively). Only when reducing the training set to 450 entries did we observe a significant discrepancy ($P=.005$ and $P=.005$, respectively).

The classifier's performance maintained an AUPRC >0.80 in settings with highly different disease prevalence (Multimedia Appendix 4). Only when disease prevalence was below 4% or above 50% did we detect a difference in performance compared to that of the initial 8% prevalence.

Cut-Off Selection

We picked the support vector machine classifier with the median performance in the training stage. This classifier assigns a probability of being a rheumatoid arthritis to each patient by summing the coefficients of the features present in the clinical notes of the patient (Figure 3). The probability cut-offs for optimized PPV (>0.95) and optimized sensitivity (>0.95) were 0.99 and 0.53, respectively (Figure 4).

The probability cut-off for optimized PPV resulted in the following test characteristics: PPV 0.96, sensitivity 0.70, specificity 1.00, negative predictive value [NPV] 1.00, and F1 score 0.81. The probability cut-off for optimized sensitivity resulted in the following test characteristics: PPV 0.72, sensitivity 0.96, specificity 0.97, NPV 1.00, and F1 score 0.82.

Figure 3. The relative importance (coefficients) of the top 20 features in the Leiden-A data set according to the final support vector machine model. The initial data was in Dutch, we translated the words to English in this figure to improve readability.

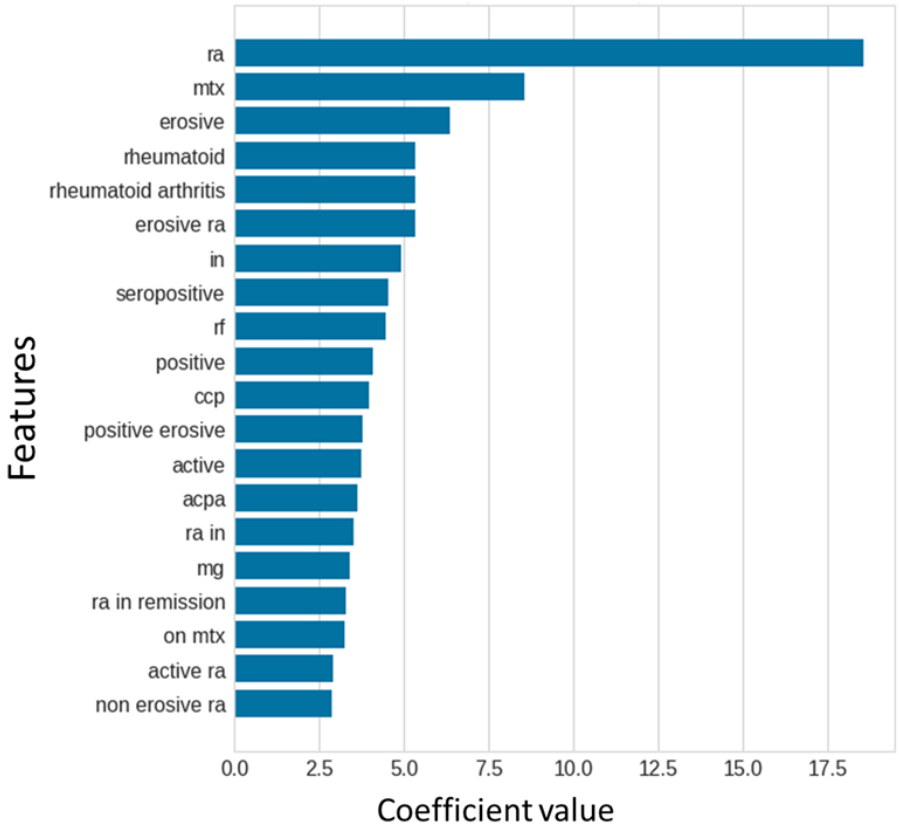
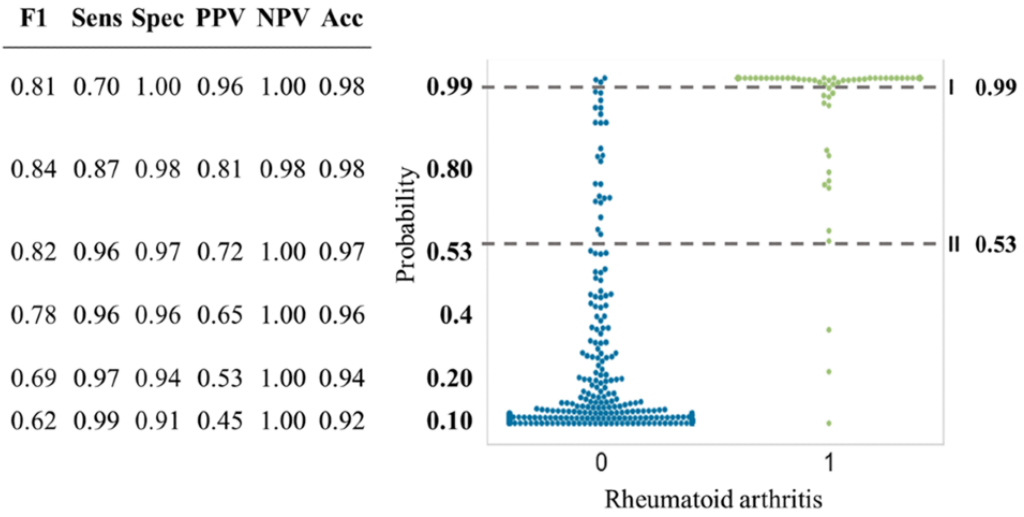


Figure 4. Swarm plot depicting the support vector machine–derived probability of being either non-rheumatoid arthritis (blue) or rheumatoid arthritis (green) for the Leiden-A data set. The dotted lines display the optimal cutoffs. Sens: sensitivity, Spec: specificity; PPV: positive predictive value; NPV: negative predictive value; Acc: accuracy; F1: F1 score.



Final Method Testing of Case Identification

In the Leiden-B data set, rheumatoid arthritis support vector machine classifier (Table 1) identified 64 cases with a cut-off of 0.99 (with corresponding PPV 0.94, sensitivity 0.71, specificity 1.00, NPV 0.97, and F1 score 0.81) and 104 cases with a cut-off of 0.53 (with corresponding PPV 0.75, sensitivity 0.93, specificity 0.97, NPV 0.99, and F1 score 0.83). In the complete Leiden data set of 23,300 patients using the first (precise) cut-off resulted in 2873 cases of rheumatoid arthritis and the second (inclusive) cut-off resulted in 6453 cases of rheumatoid arthritis.

Table 1. Support vector machine confusion matrices for the Leiden-B test set (n=1000) .

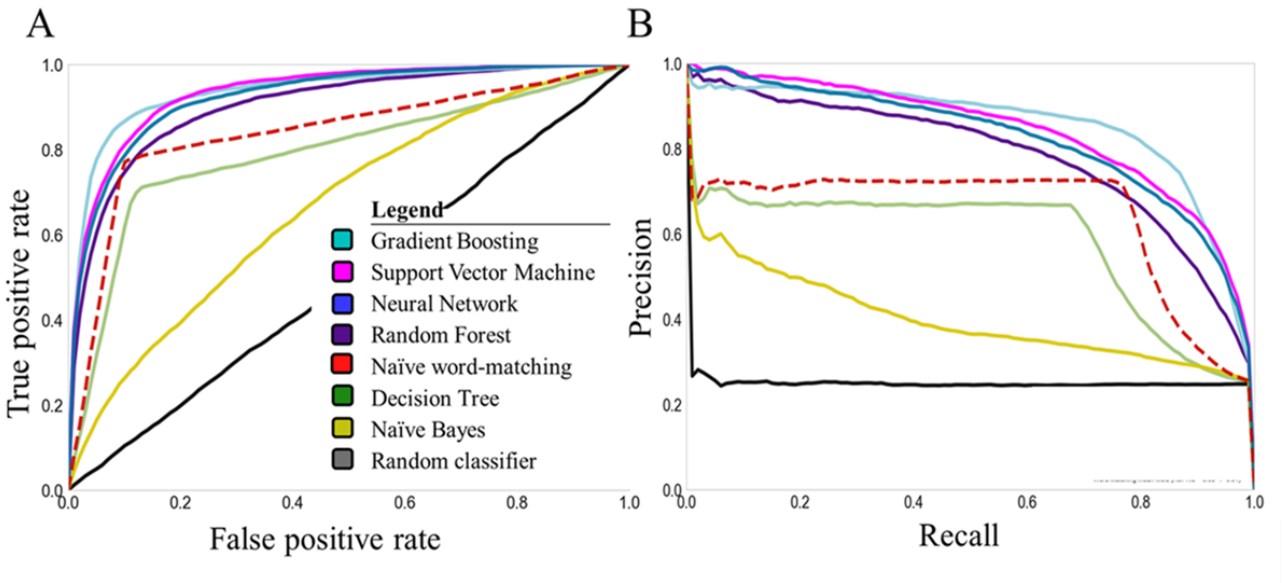
Clinician-based	Support vector machine 1 (cut-off=0.99)		Support vector machine 2 (cut-off=0.53)	
	Non–rheumatoid arthritis	Rheumatoid arthritis	Non–rheumatoid arthritis	Rheumatoid arthritis
Non–rheumatoid arthritis	912 (true negative)	4 (false positive)	890 (true negative)	26 (false positive)
Rheumatoid arthritis	24 (false negative)	60 (true positive)	6 (false negative)	78 (true positive)

Validation of Workflow in Erlangen Data

Training and Model Selection

To evaluate the universal applicability of the workflow, we employed the full pipeline on Erlangen data sets. Again, we ran all machine learning methods to find the best performing method using the Erlangen-A data set. Gradient boosting achieved the best performance (AUROC 0.94; AUPRC 0.85; F1 score 0.81) (Figure 5). The probability cut-offs for optimized PPV (> 0.90) and optimized sensitivity (>0.90) were 0.79 and 0.19, respectively (Multimedia Appendix 5).

Figure 5. (A) Receiver operating characteristics and (B) precision-recall curves for all machine learning methods (solid lines) and the naïve word-matching method (dotted line) in the training set (Erlangen-A).



Final Method Testing of Case Identification

When we applied the model on the test data set (Erlangen-B), we obtained similar performance (Table 2) with the predefined cut-offs as those found for the training data set (Erlangen-A). The gradient boosting classifier identified 59 cases with a cut-off of 0.79 (with corresponding PPV 0.97, sensitivity 0.51, specificity 0.99, NPV 0.87, and F1 score 0.67) and 131 cases with the cut-off of 0.19 (with corresponding PPV 0.72, sensitivity 0.84, specificity 0.90, NPV 0.95, and F1 score 0.77).

Table 2. Gradient boosting confusion matrices for the Erlangen-B test set (n=478).

Clinician-based	Gradient boosting 1 (cut-off=0.79)		Gradient boosting 2 (cut-off=0.19)	
	Non–rheumatoid arthritis	Rheumatoid arthritis	Non–rheumatoid arthritis	Rheumatoid arthritis
Non–rheumatoid arthritis	364 (true negative)	2 (false positive)	329 (true negative)	37 (false positive)
Rheumatoid arthritis	55 (false negative)	57 (true positive)	18 (false negative)	94 (true positive)

Discussion

Principal Findings

Our study describes the results of a pipeline that applies multiple machine learning methods as well as naïve word-matching to create algorithms of case selection (patients with rheumatoid arthritis in our example) from electronical medical records. We observed that most methods outperform a naïve word matching algorithm. Our pipeline created algorithms on both Dutch and German data that showed a high performance in the testing and validation phase (F1 score 0.83 and 0.82 respectively). When we defined the cut-offs for case selection from the first data set aiming for either a high sensitivity or high PPV, we observed that the performances were robust in the second data sets (Leiden-B: PPV 0.94 and sensitivity 0.93; Erlangen-B: PPV 0.97 and sensitivity 0.84).

We believe that our approach of making a center-specific algorithm is more attractive than the application of an algorithm developed elsewhere, since our method is more precise, doesn't require standardization, and most importantly, it ensures high

performance within the center. Our method only requires similar effort as the application of predefined algorithms, namely chart reviewing a subset of data. Furthermore, our workflow respects the user's requirements regarding the case selection. The case selection can be tailored to being highly precise or sensitive depending on the chosen cut-off.

Furthermore, this study shows the power of machine learning approaches to generate cohorts of patients in seconds, laying a foundation for allowing studies of cohorts with an unprecedented low cost.

When applying our support vector machine classifier on the complete Leiden University Medical Centre's database of 23,300 cases (including the 3000 annotated records) we identified 2873 rheumatoid arthritis cases when employing the stringent probability threshold of 0.99. The automatic annotation only took 6.17 seconds, a fraction of the amount of time it would take to review the medical charts manually.

Future Directions

Our aim was to implement a broadly applicable workflow. The current versions require installing Anaconda (version 5.1.0) and Python (version 3.6). Researchers without any computational experience might feel certain reluctance to start the pipeline. We tested (without quantification) how easy someone outside our center could run the pipeline, by sending the scripts to scientists at Erlangen. Though they implemented the pipeline with relative ease, we do acknowledge that it was done by someone with experience in computational languages. Also, testing the pipeline in Erlangen exposed some unclarities in the scripts, which have been improved. The next step would be to perform a usability study, where we could ask users for their experience as well as test how much time it takes them to get the script running. We could further improve the usability of the pipeline by creation of a web-based interface where people could upload their data and get back their results automatically. This would require substantial computational resources as the data sets are large. In addition, we would need to ensure encryptions processes as clinical notes have a high risk to breach privacy.

Limitations

We want to note 3 important shortcomings of our study. The first limitation is that deploying the pipeline requires user familiarity with implementation software. Our proposed workflow facilitates building a classifier with a step-by-step implementation. Affinity with programming is not required, because all functions for training and evaluation are already provided. However, some software experience is beneficial when setting up the environment for the pipeline to run. With the emergence of machine learning and natural language processing we would argue that it becomes increasingly useful to possess the skills required to implement software.

Second, we acknowledge that the workflow was evaluated in only 2 centers, both with Germanic languages. Although the pipeline provides language-specific preprocessing with pretrained tools for most languages, it would be interesting to investigate if similar performance can be achieved in centers with low lexical similarities to the Dutch language (eg, languages without a Latin-based alphabet).

Finally, we acknowledge that the models' performances can be further optimized by fine-tuning hyperparameters. These are parameters of the machine learning method that are provided prior to training the machine learning method. Additionally, it is possible to adjust the size of the n-grams to improve the performance. Since our models consistently performed very well in training and testing, we did not optimize any parameters in our study. Furthermore, we only evaluated a handful of candidate machine learning methods. Our selection is by no means an exhaustive list of available techniques in the field. We selected these methods as they cover a variety of machine learning method and are widely known.

Lessons Learned

We were able to conduct a stringent flow of training and testing, whereby we used several independent data sets to, first, optimize the classifiers, and second, to ensure reliable calculations of the classifiers' performances by using k-fold cross-validation and both receiver operating characteristic and precision recall curves on 10-fold cross-validation, providing a good indication of performance on unseen data.

To select the best classifier, we performed paired *t* tests on 5×2 cross-validation rather than 10-fold cross-validation. Although performing a paired *t* test on 10-fold cross-validation is a very common practice, we learned that this test is not recommended. The correlation between overlaps violates the *t* test's assumption of independence, resulting in more false positives (increased type I error); 5×2 cross-validation splits the data set 50/50 and is, therefore, more suitable for statistical analysis. However, 5×2 cross-validation is confined to a small training set, which is why we also used 10 cross-validations to approximate the performance on unseen data.

Our study is not the first to examine methods for disease identification from EHR [3]. Studies have employed high-throughput methods on structured data such as ICD (billing) codes. Regrettably, such codes have a poor performance because they describe why a patient is examined, which does not strictly mean that a patient has that diagnosis. More successful algorithms (often called phenotype algorithms) combined a variety of methods including rule-based case identification and natural language processing [2]. Though these algorithms have a good median performance when tested in multiple clinics, on an individual center PPV varies (below 0.5 for several clinics) [2]. Moreover, several centers required additional tailoring to allow application of the algorithms. This is not surprising since health clinics have different protocols for registering information.

As gold standard, we purposely chose the diagnosis of the treating rheumatologist in contrast to counting the disease classification criteria [17,18]. The problem with the latter is that classification criteria have been developed for research and not for clinical practice where all information including additional tests in the differential diagnostic workup are taken into account. Moreover, the exact information for individual criteria is often not precisely registered in EHRs.

We ran several sensitivity analyses to explore the influence of disease prevalence and number of selected patients on the model's performance. The support vector machine classifier was robust over different selections of training data (low standard error on the cross-validation results), number of training samples, and imbalances of case number. These analyses also showed that in our Leiden data the annotation of 600 patients would have been sufficient to build a reliable classifier. We acknowledge that due to difference in feature variance, the optimal number of patients required to train the classifier might differ between centers.

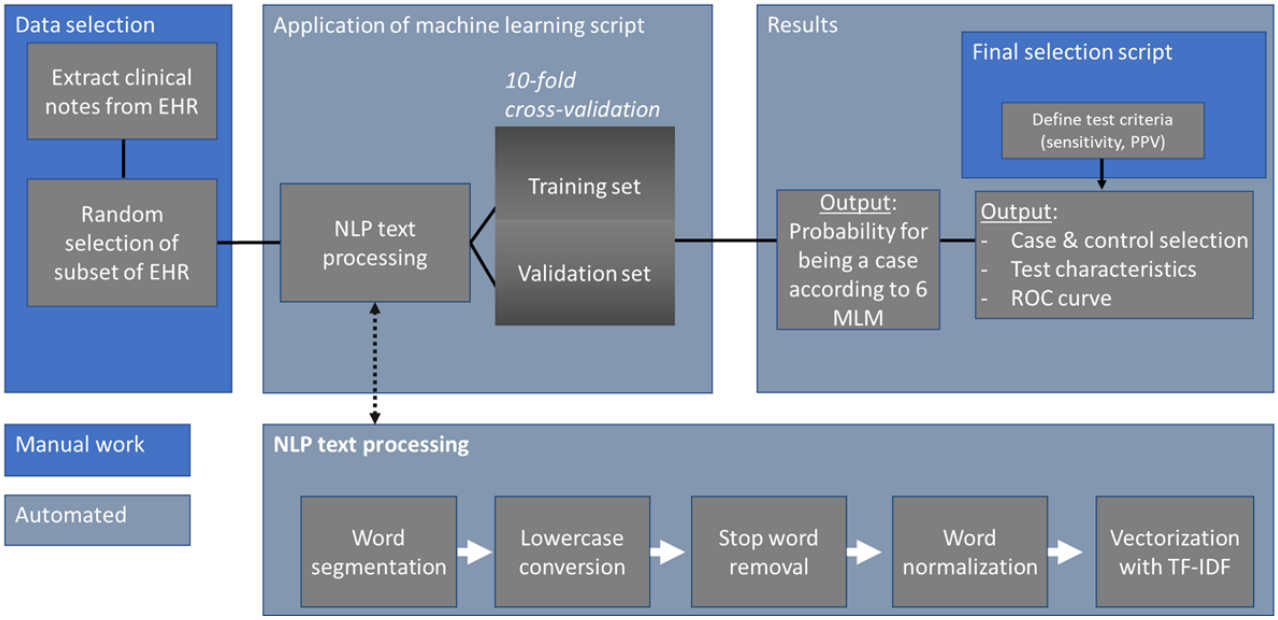
Generalizability of the Workflow

The support vector machine was the best classifier for Leiden-A (F1 score 0.83), although the difference was not significant with respect to the gradient boosting, neural networks, and random forest. The support vector machine was employed in the independent Leiden-B data set with similarly good performance (F1 score 0.81). We predefined 2 thresholds of the rheumatoid arthritis support vector machine probabilities on the first Leiden data (Leiden-A) aiming for either a high precision (PPV 0.94), or a high sensitivity (sensitivity 0.93). When we applied these predefined cut-offs in the second set of patients we obtained similarly high test characteristics (PPV 0.96, sensitivity 0.70, specificity 1.00, NPV 1.00 with the highly precise threshold, and PPV 0.72, sensitivity 0.96, specificity 0.97, NPV 1.00 with the highly sensitive threshold). Finally, we ran the same workflow of training and testing as employed on the Dutch Leiden data to the German Erlangen data. Again, we built a high performing classifier (in this case gradient boosting performed best) that gave consistent results for both settings (PPV 0.97, sensitivity 0.51, specificity 0.99, NPV 0.87 with the highly precise threshold, and PPV 0.72, sensitivity 0.84, specificity 0.90, NPV 0.95 with the highly inclusive threshold).

The gradient boosting has the best performance in the Erlangen data, while in the Leiden data the support vector machine performs the best. This is not necessarily surprising, as “there is no such thing as a free lunch” (meaning that a universal best algorithm does not exist) [19]. The high performance of the support vector machine is achieved by generalizing the Leiden data. There is no guarantee that the technique used in the Leiden data set will also perform the best in the Erlangen data set. Notably, in each data set, both methods performed very well with only very modest differences. The slight deviations in performance between the methods could be caused by language differences and characteristic notations of the center.

In accordance with the FAIR principles [20], we have made all our scripts publicly available and optimized them so scientists may use them regardless of prior experience (Figure 6) [16]. We advise centers not to use our specific classifier but to follow the workflow as presented in this paper and build a classifier that fits the local data best.

Figure 6. Flowchart describing the steps to apply the machine learning scripts to new data. EHR: electronic health record; MLM: machine learning method; NLP: natural language processing; PPV: positive predictive value; ROC: receiver operating characteristic; TF-IDF: term frequency by inverse document frequency.



Conclusion

The workflow facilitates the production of highly reliable center-specific machine learning methods for the identification of patients with rheumatoid arthritis from format-free text fields. Our results suggest that our workflow can easily be applied to other EHRs or other diseases and is not restrained by specific language, EHR software, or treatments. This methodology of machine learning for EHR data extraction facilitates cohort studies (with regard to cost and size).

Acknowledgments

This study was supported by the Dutch Arthritis Association (*ReumaNederland*) 15-3-301 and by Measurement of Efficacy of Treatment in the 'Era of Outcome' in Rheumatology (project number RP 2014-03).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of study details.
[DOCX File , 20 KB - medinform_v8i11e23930_app1.docx]

Multimedia Appendix 2
Average F1 score for all machine learning methods (Leiden-A).
[PNG File , 91 KB - medinform_v8i11e23930_app2.png]

Multimedia Appendix 3
Performance of SVM on increasing training set within the Leiden-A data set.
[PNG File , 442 KB - medinform_v8i11e23930_app3.png]

Multimedia Appendix 4
Performance of SVM on increasing prevalence in Leiden-A.
[PNG File , 256 KB - medinform_v8i11e23930_app4.png]

Multimedia Appendix 5
Swarm plot depicting gradient boosting–derived probability of being rheumatoid arthritis.
[PNG File , 315 KB - medinform_v8i11e23930_app5.png]

References

1. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010 Aug;62(8):1120-1127 [FREE Full text] [doi: 10.1002/acr.20184] [Medline: 20235204]
2. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016 Nov;23(6):1046-1052 [FREE Full text] [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
3. Jamian L, Wheless L, Crofford L, Barnado A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. Arthritis Res Ther 2019 Dec 30;21(1):305 [FREE Full text] [doi: 10.1186/s13075-019-2092-7] [Medline: 31888720]
4. van den Berg R, van der Heijde D, Landewé R, van Lambalgen K, Huizinga T. The METEOR initiative: the way forward for optimal, worldwide data integration to improve care for RA patients. Clin Exp Rheumatol 2014;32(5 Suppl 85):S-135. [Medline: 25365103]
5. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
6. Manning CD, Raghavan P, Schutze H. Text classification and naive Bayes. In: Introduction to Information Retrieval. Cambridge: Cambridge University Press; Jul 7, 2008:253-289.
7. Lightbody G, Irwin GW. Multi-layer perceptron based modelling of nonlinear systems. Fuzzy Sets and Systems 1996 Apr 8;79(1):93-112. [doi: 10.1016/0165-0114(95)00293-6]
8. Breiman L. Random forests. Machine Learning 2001;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324] [Medline: 21816105]
9. Bennett KP, Campbell C. Support vector machines. SIGKDD Explor Newsl 2000 Dec 01;2(2):1-13. [doi: 10.1145/380995.380999]
10. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001 Oct;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
11. Uysal AK, Gunal S. The impact of preprocessing on text classification. Information Processing & Management 2014 Jan;50(1):104-112. [doi: 10.1016/j.ipm.2013.08.006]
12. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 2010 Mar;32(3):569-575. [doi: 10.1109/tpami.2009.187]

13. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
14. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998 Sep 15;10(7):1895-1923. [doi: 10.1162/089976698300017197] [Medline: 9744903]
15. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management 2009 Jul;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
16. Maarseveen T. DiagnosisExtraction_ML. Github. URL: https://github.com/levrex/DiagnosisExtraction_ML [accessed 2020-10-01]
17. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988 Mar;31(3):315-324. [doi: 10.1002/art.1780310302] [Medline: 3358796]
18. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, et al. 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010 Sep;62(9):2569-2581 [FREE Full text] [doi: 10.1002/art.27584] [Medline: 20872595]
19. Wolpert DH, Macready W. No free lunch theorems for optimization. IEEE Trans Evol Computat 1997 May;1(1):67-82. [doi: 10.1109/4235.585893]
20. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(160018) [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]

Abbreviations

AUPRC: area under the precision recall curve
AUROC: area under the receiver operating characteristic curve
EHR: electronic health record
ICD-9: International Classification of Diseases, Ninth Revision
NPV: negative predictive value
PPV: positive predictive value

Edited by C Lovis; submitted 28.08.20; peer-reviewed by L Rodriguez Rodriguez, C Popa; comments to author 20.09.20; revised version received 18.10.20; accepted 24.10.20; published 30.11.20.

Please cite as:
Maarseveen TD, Meinderink T, Reinders MJT, Knitza J, Huizinga TWJ, Kleyer A, Simon D, van den Akker EB, Knevel R
Machine Learning Electronic Health Record Identification of Patients with Rheumatoid Arthritis: Algorithm Pipeline Development and Validation Study
JMIR Med Inform 2020;8(11):e23930
URL: http://medinform.jmir.org/2020/11/e23930/
doi:10.2196/23930
PMID:33252349

Corrigenda and Addenda

Correction: Use of an Electronic Clinical Decision Support System in Primary Care to Assess Inappropriate Polypharmacy in Young Seniors With Multimorbidity: Observational, Descriptive, Cross-Sectional Study

Eloisa Rogero-Blanco[1,2,3*], MD; Juan A Lopez-Rodriguez[1,2,3,4*], MD, MSc, PhD; Teresa Sanz-Cuesta[3,4*], MD, PhD; Mercedes Aza-Pascual-Salcedo[5*], PharmD; M Jose Bujalance-Zafra[6*], MD, PhD; Isabel Cura-Gonzalez[2,3,4*], MD, MPH, PhD; MultiPAP Group[7*]

[1]General Ricardos Primary Health Care Centre, Madrid, Spain

[2]Medical Specialties and Public Health, School of Health Sciences, University Rey Juan Carlos, Alcorcón, Madrid, Spain

[3]Health Services Research on Chronic Patients Network (REDISSEC), Madrid, Spain

[4]Research Support Unit, Primary Care Management, Madrid, Spain

[5]Dirección Farmacia Atención Primaria Sector Zaragoza III, Zaragoza, Spain

[6]Dirección Unidad Gestión Clínica Victoria en Málaga, Servicio Andaluz de Salud, Málaga, Spain

[7]See Acknowledgments

[*]all authors contributed equally

Corresponding Author:
Juan A Lopez-Rodriguez, MD, MSc, PhD
General Ricardos Primary Health Care Centre
Calle General Ricardos
Madrid, 28019
Spain
Phone: 34 685197913
Email: juanantonio.lopez@salud.madrid.org

Related Article:

Correction of: https://medinform.jmir.org/2020/3/e14130

(JMIR Med Inform 2020;8(11):e25678) doi:10.2196/25678

In "Use of an Electronic Clinical Decision Support System in Primary Care to Assess Inappropriate Polypharmacy in Young Seniors With Multimorbidity: Observational, Descriptive, Cross-Sectional Study" (JMIR Med Inform 2020;8(3):e14130) the authors noted one issue.

In the Acknowledgments section of the paper, the statement:

> *This study was financed by the Fondo de Investigaciones Sanitarias Coordinated project (grant references PI15/00276, PI15/00572, PI15/00996) from the Instituto de Salud Carlos III*

has been removed and replaced by the following statement:

> *This study was funded by National Institute for Health Research ISCIII (Grant numbers PI15/00276 (APT), PI15/00572 (ICG), PI15/00996 (JDPT), RD16/0001/0004 (ICG), RD16/0001/0005 (APT), RD16/0001/0006 (JDPT)) Co-funded by European Regional Development Fund, (ERDF) "A way of shaping Europe". National Plan I+D+I 2013-2016.*

The correction will appear in the online version of the paper on the JMIR Publications website on November 19, 2020, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 11.11.20; this is a non–peer-reviewed article;accepted 11.11.20; published 19.11.20.

Please cite as:
Rogero-Blanco E, Lopez-Rodriguez JA, Sanz-Cuesta T, Aza-Pascual-Salcedo M, Bujalance-Zafra MJ, Cura-Gonzalez I, MultiPAP Group
Correction: Use of an Electronic Clinical Decision Support System in Primary Care to Assess Inappropriate Polypharmacy in Young Seniors With Multimorbidity: Observational, Descriptive, Cross-Sectional Study
JMIR Med Inform 2020;8(11):e25678
URL: http://medinform.jmir.org/2020/11/e25678/
doi:10.2196/25678
PMID:33211669

Original Paper

Universal Patient Identifier and Interoperability for Detection of Serious Drug Interactions: Retrospective Study

Howard Michael Sragow[1], MBA; Eileen Bidell[1], RPH; Douglas Mager[2], MA; Shaun Grannis[3], MD

[1]Express Scripts Inc, Franklin Lakes, NJ, United States

[2]Express Scripts Inc, St. Louis, MO, United States

[3]Regenstrief Institute, Indiana University Medical School, Indianapolis, IN, United States

Corresponding Author:
Howard Michael Sragow, MBA
Express Scripts Inc
100 Parsons Pond Drive
Franklin Lakes, NJ, 07417
United States
Phone: 1 201 269 4342
Email: howard_sragow@express-scripts.com

Abstract

Background: The United States, unlike other high-income countries, currently has no national unique patient identifier to facilitate health information exchange. Because of security and privacy concerns, Congress, in 1998, prevented the government from promulgating a unique patient identifier. The Health and Human Services funding bill that was enacted in 2019 requires that Health and Human Services report their recommendations on patient identification to Congress. While there are anecdotes of incomplete health care data due to patient misidentification, to date there have been insufficient large-scale analyses measuring improvements to patient care that a unique patient identifier might provide. This lack of measurement has made it difficult for policymakers to balance security and privacy concerns against the value of potential improvements.

Objective: We sought to determine the frequency of serious drug-drug interaction alerts discovered because a pharmacy benefits manager uses a universal patient identifier and estimate undiscovered serious drug-drug interactions because pharmacy benefit managers do not yet fully share patient records.

Methods: We conducted a retrospective study of serious drug-drug interaction alerts provided from September 1, 2016 to August 31, 2019 to retail pharmacies by a national pharmacy benefit manager that uses a unique patient identifier. We compared each alert to the contributing prescription and determined whether the unique patient identifier was necessary in order to identify the crossover alert. We classified each alert's disposition as override, abandonment, or replacement. Using the crossover alert rate and sample population size, we inferred a rate of missing serious drug-drug interaction alerts for the United States. We performed logistic regression in order to identify factors correlated with crossover and alert outcomes.

Results: Among a population of 49.7 million patients, 242,646 serious drug-drug interaction alerts occurred in 3 years. Of these, 2388 (1.0%) crossed insurance and were discovered because the pharmacy benefit manager used a unique patient identifier. We estimate that up to 10% of serious drug-drug alerts in the United States go undetected by pharmacy benefit managers because of unexchanged information or pharmacy benefit managers that do not use a unique patient identifier. These information gaps may contribute, annually, to up to 6000 patients in the United States receiving a contraindicated medication.

Conclusions: Comprehensive patient identification across disparate data sources can help protect patients from serious drug-drug interactions. To better safeguard patients, providers should (1) adopt a comprehensive patient identification strategy and (2) share patient prescription history to improve clinical decision support.

(JMIR Med Inform 2020;8(11):e23353) doi:10.2196/23353

KEYWORDS

patient identification; pharmacy benefit manager; interoperability; adverse drug event; identity management; identifier; pharmacy; pharmaceuticals; drug

Introduction

Patient Identification in the United States

Interoperability is a key factor in the quality of health care [1-3]. Many anecdotes describe information failing to reach a provider, or providers overlooking records belonging to the same patient, hindering clinical decision making [4,5]. The Centers for Medicare and Medicaid Services Interoperability and Patient Access final rule [6] facilitates better exchange, but without consistent patient identification, its success will be limited. Comprehensive patient identification accurately and efficiently integrates typically fragmented patient data to create a more complete record while mitigating the incorrect linkage of health care data belonging to other patients.

American providers currently do not have a national unique patient identifier to facilitate patient information exchange. Congress, in 1998, prevented the government from promulgating a unique patient identifier by prohibiting funding for such an initiative. Social security number use was not explicitly prohibited, and its use in health care continues. However, privacy concerns persist, and use of social security numbers for health care identity management is steadily declining [7].

The unique patient identifier debate continued until 2019, when the House and Senate bills funding the Department of Health and Human Services diverged. The House bill [8,9] would have enabled the Department of Health and Human Services to promulgate a unique patient identifier; the Senate bill [10] would not. The law enacted in December 2019 was a compromise, requiring that the Department of Health and Human Services report recommendations on patient identification to Congress [11,12]. This study is submitted in part to help inform that recommendation.

While differing models for unique patient identifier assignment exist, a prevailing model in many high-income countries leverages entry events. A central system recognizes an event that occurs once per patient (eg, birth, immigration) and assigns a unique patient identifier. Providers then use the centrally assigned unique patient identifier to identify the patient.

Another model that is common in US health care systems uses demographic matching to assign a unique patient identifier. Providers use demographic information (eg, first name, last name, date of birth, address) to identify the patient, applying the existing unique patient identifier if successfully identified. Otherwise, a new unique patient identifier is assigned. Demographic matching is susceptible to error and as demographics change, providers risk incorrectly duplicating or merging patients.

Patient Identification and Interoperability in Pharmacy Benefit Managers

Pharmacy is one area of health care where inexact identification can adversely impact patients. Prescription history enables providers to help patients avoid serious drug-drug interactions. Although estimates for serious drug-drug interaction risks vary [13-17], there is ample evidence that they can be dangerous.

To mitigate these risks, both pharmacy benefit managers and dispensing pharmacies perform prospective drug utilization review using prescription history before dispensing drugs [18,19], which assesses the requested medication in the context of the patient's prescription history and is well-established in pharmacy practice. When electronic review identifies a potentially serious drug-drug interaction, the pharmacy benefit manager alerts the pharmacist through a claim rejection.

Although pharmacy benefit managers process two-thirds of prescriptions in the United States [20], they may lack access to comprehensive prescription histories. During claim adjudication, pharmacy benefit managers aggregate prescriptions filled by multiple pharmacies, creating a history that is more complete than that of any single pharmacy. While pharmacy benefit manager intervention is a secondary defense against serious drug-drug interactions, it augments other medication safeguards.

However, pharmacies can capture prescription history information to which pharmacy benefit managers lack access. Patients may self-pay or obtain reimbursement of prescription costs through manufacturer coupons. Complete visibility into a patient's prescription history is also limited when a patient transitions between pharmacy benefit managers.

If the dispensing pharmacy system lacks the patient's prior prescriptions, a labor-intensive process to obtain prescription history may be needed, rendering automated prospective drug utilization review less effective. Consequently, without complete electronic prescription data, automated prospective drug utilization reviews fail to adequately detect potentially serious problems.

Some pharmacy benefit managers have technology to detect records with similar demographics, assigning those to a single unique patient identifier, and using that unique patient identifier during prospective drug utilization review. Other pharmacy benefit managers may use a beneficiary identifier to identify the patient. The latter would treat a record from a new payor as a new patient, omitting the relevant prescription history from prospective drug utilization review and missing serious drug-drug interactions.

Patients risk serious drug-drug interactions going unidentified when their beneficiary identifier changes or when a different pharmacy benefit manager assumes management of their prescriptions. Changes can occur in 1 of 3 ways: (1) The patient changes payor (eg, insurer, employer, labor union, etc) and the new payor uses a different pharmacy benefit manager. Upon a benefit change (employment change, Medicare eligibility, work injury, discount card usage, etc), if the new payor uses a different pharmacy benefit manager, that new pharmacy benefit manager typically does not obtain the patient's prescription history from the prior pharmacy benefit manager. (2) The payor chooses a new pharmacy benefit manager. When payors select a different pharmacy benefit manager, prescription histories are not always forwarded to the new pharmacy benefit manager. (3) The patient changes payor, but the new and old payors happen to use the same pharmacy benefit manager. The patient adopts a new benefit but keeps the same pharmacy benefit manager.

Objective

To date, few formal studies have evaluated how interoperability or use of a unique patient identifier affects clinical results. While prior studies have examined potential cost savings [21,22], we are unaware of studies assessing the relationship between use of a unique patient identifier or interoperability and clinical outcomes for large populations.

Regardless of the strategy used to assign a unique patient identifier, increasing evidence links inaccurate identification to poor outcomes [23-28]. Thus, some health care identity experts opine that improved identification methods such as a unique patient identifier may reduce adverse patient outcomes [29]. This study compares the edits resulting from application of a unique patient identifier to prospective drug utilization review with those from not using a unique patient identifier to measure the improvements a unique patient identifier might provide and forecast improvements resulting from broad interoperability.

We hypothesize that using a unique patient identifier to aggregate prescription history for prospective drug utilization review can improve the completeness of patient prescription history, yielding more accurate detection of drug-drug interactions. Our primary objective was to quantify the extent to which a unique patient identifier can improve prospective drug utilization review's ability to identify serious drug-drug interactions compared to that when using a beneficiary identifier. To contextualize the rate of missing alerts, we measured how often patients migrate between pharmacy benefit managers.

Our second objective was to forecast the improvement in prospective drug utilization review accuracy under the assumption that clinical decision makers have comprehensive access to patient prescription history enabled by a broadly available unique patient identifier.

Methods

This retrospective analysis uses serious drug-drug interaction alerts that were provided to retail pharmacies at the time of adjudication from September 2016 to August 2019 among a patient population of 49.7 million serviced by a large national pharmacy benefit manager. The pharmacy benefit manager aggregates prescription history data for real-time prospective drug utilization review using proprietary deterministic algorithms to link records to the same unique patient identifier. While no algorithm is perfect, a unique patient identifier can improve patient record completeness [30].

Drug interaction alerts are triggered by a prescription and a precipitating claim. For our first objective, we determined whether each serious drug-drug interaction alert was captured using the same health insurance identifier for both the prescription and precipitating claim. When the precipitating claim originated under different insurance in the absence of a unique patient identifier, we assumed that the pharmacy benefit manager failed to detect serious drug-drug interaction and generated no alert. When the unique patient identifier identified the precipitating claim, enabling prospective drug utilization review to trigger a serious drug-drug interaction alert despite different insurance, we called this event a *crossover alert*. We measured how often crossover alerts occur, relative to all alerts.

For our second objective, we categorized each serious drug-drug interaction alert into 1 of 3 outcomes:

1. Override: The patient receives the medication subsequent to internal pharmacy review within 14 days of the serious drug-drug interaction alert.
2. Replacement: The patient receives another medication treating the same condition within 14 days.
3. Abandonment: The patient did not receive a prescription for another medication treating that condition within 14 days.

We performed chi-square and student *t* tests on bivariate findings and used logistic regression to identify factors correlated with crossover alerts and each of the 3 outcomes (abandonment, override, and replacement). We examine covariates with nonscalar data separated into dichotomous factors representing the more commonly occurring values, including drug, by First Databank specific therapeutic class; Drug Enforcement Agency schedule; month and year of service; type of pharmacy benefit, including Medicare Part D; Exchange Plan under the Affordable Care Act; Medicaid; and Other, including commercial and employer plans; patient age and gender as reported by the payor. We present odds ratios from the logistic regressions alongside the bivariate findings in each of the specific results sections. We also performed a multinomial logistic regression for abandonment and replacement compared to override, in order to rule out inflation of any significance measurements.

Using the serious drug-drug interaction crossover alerts observed and the market share of the pharmacy benefit manager population studied and applying national proportional weights for gender and age distributions, we estimated rates of missing alerts for the entire US insured population.

We assumed that patients randomly remain or transition from their pharmacy benefit manager each year. New health insurance identifiers are typically assigned not by the pharmacy benefit manager but by the payor. Except for patients choosing a new Medicare Part D plan, a patient does not directly select a pharmacy benefit manager, and patients do not choose a new employer on the basis of the pharmacy benefit manager serving the employees. We also assumed that other factors influencing serious drug-drug interactions (ie, demographics, prescribing patterns, self-pay rates, etc) in the observed and unobserved populations were similar.

Crossover alerts were observed in a subset of the US population. We assumed that if pharmacy benefit managers could access prescription records for the remaining population using a common unique patient identifier, crossover alerts of serious drug-drug interactions would reflect all transitioning patients, rather than just those transitioning within a pharmacy benefit manager. Alerts would increase by a proportion that we labeled a *proportionality factor*, which was defined as the US insured population (91.5% of 328 million) [31] divided by the population studied. We again used age and gender weightings to estimate annual serious drug-drug interaction alerts for the entire US population. We estimated an additional crossover

alert percentage by multiplying the crossover alert percentage by the proportionality factor.

While serious drug-drug interaction alerts resulting in replacement or abandonment can improve outcomes, we assumed that overridden alerts do not. To estimate annual unidentified alerts that might have helped prevent a contraindicated dispensing, we counted only the proportion that would have resulted in abandonment or replacement as the product of (1) annual national alerts, (2) the crossover alert percentage, and (3) the percentage of serious drug-drug interaction alerts abandoned or replaced.

To contextualize the rate of missing alerts, we measured how often patients migrate between pharmacy benefit managers independently of payor transition. We identified patients taking 2 commonly prescribed medications with indications for long-term preventative therapy for highly prevalent chronic conditions: atorvastatin, indicated for hyperlipidemia, and amlodipine besylate, for hypertension. We expected a high proportion of these patients would have regular claims for these medications throughout the study period. We select patients aged 38 to 48 years old, a range associated with a 2-year mortality rate lower than 1% [32] and identified 2 cohorts: one with at least 2 claims for 1 or both of these medications in 2017, and one with at least 2 claims in the year 2019. For the 2019 cohort, we determined the proportion of patients that were present in 2018 and 2017. For the 2017 cohort, we measured yearly attrition rates in 2018 and 2019.

Results

General

For the 49.7 million patients (16.5% of insured population) included in the analysis, 1,436,799,263 total claims were processed during the study period.

From those claims, prospective drug utilization review identified 242,646 serious drug-drug interaction alerts. Among those alerts, 2388 (0.98%) were crossover alerts. Consequently, approximately 1% of all serious drug-drug interaction alerts would not have been detected, were the prospective drug utilization review limited to histories linked to the patient health insurance identifier. Since 16.5% of the insured population had 242,646 serious drug-drug interaction alerts in 3 years, we estimated the US insured population has 458,285 annual serious drug-drug interaction alerts (age- and gender-adjusted).

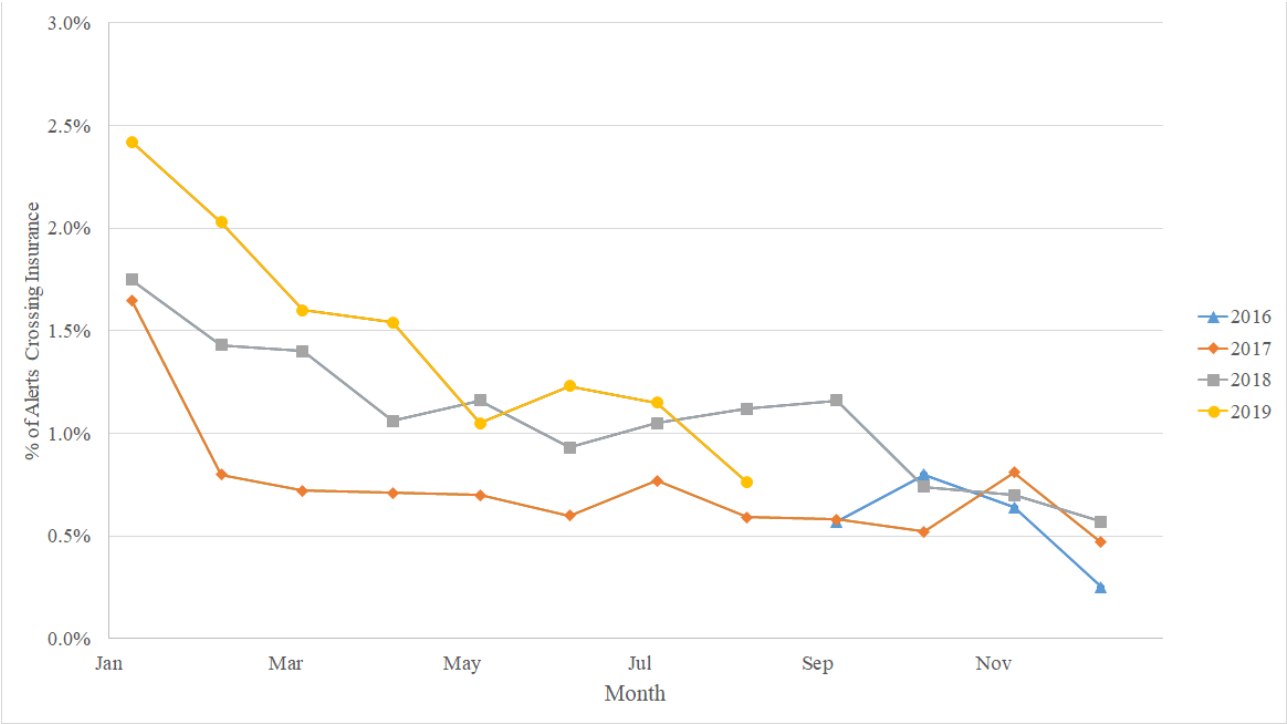
Alert Results

Of the 242,646 serious drug-drug interactions, 16.5% (40,128) were abandoned, 73.5% (178,239) were overridden, and 10.0% (24,279) were replaced. Crossover alerts were overridden and abandoned at rates indistinct from those of noncrossover alerts (noncrossover abandoned: 39,601/240,258, 16.5%, crossover abandoned: 527/2388, 22.1%; $P<.001$; noncrossover overridden: 176,551/240,258, 73.5%, crossover overridden: 1688/2388, 70.7%, $P=.002$). Significantly fewer crossover alerts were replaced compared with noncrossover alerts (173/2388, 7.2% vs 24,106/240,258, 10.0%, $P<.001$).

Month

Figure 1 shows crossover alerts occurred significantly more often in January (414/21,801, 1.9%; $P<.001$), February (276/20,226, 1.4%, $P<.001$) and March (262/21,670, 1.2%, $P<.001$), while no differences were noted for the remainder of the year. Multivariate analysis showed that alerts in January were 2.44 (95% CI 2.18-2.74) times more likely to be crossover alerts, those in February were 1.75 (95% CI 1.53-2.00) times more likely to be crossover alerts, and those in March were 1.52 (95% CI 1.33-1.74) times more likely to be crossover ($P<.001$).

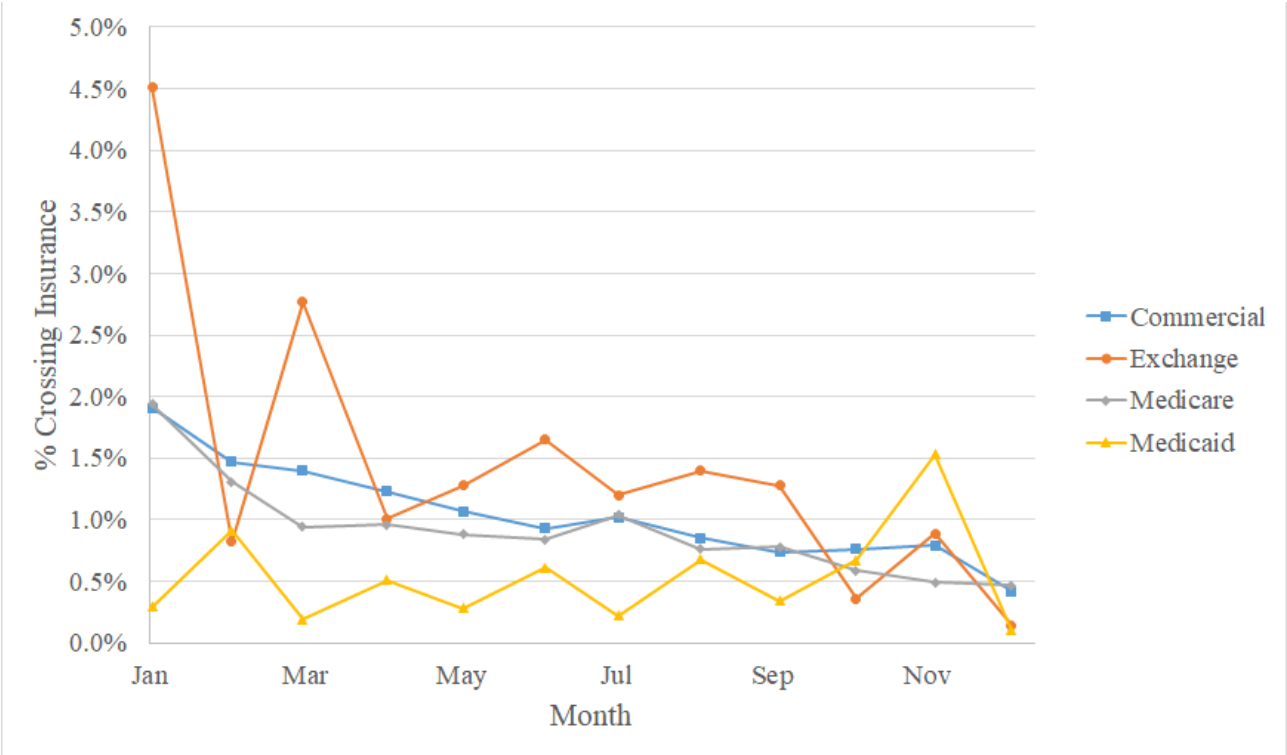
Figure 1. Monthly percentage of alerts crossing insurance identifier (September 1, 2016 to August 31, 2019).



Benefit Type

Figure 2 indicates that among Medicaid beneficiaries, crossover alerts occurred less often (60/11,668, 0.5%; *P*<.001) and did not spike in the first quarter. The percentage of serious drug-drug interaction crossover alerts among patients enrolled in an Affordable Care Act Exchange Plan was higher (85/6173, 1.4%; *P*=.002).

Figure 2. Monthly percentage of alerts crossing insurance identifier by payor type (September 1, 2016 to August 31, 2019).

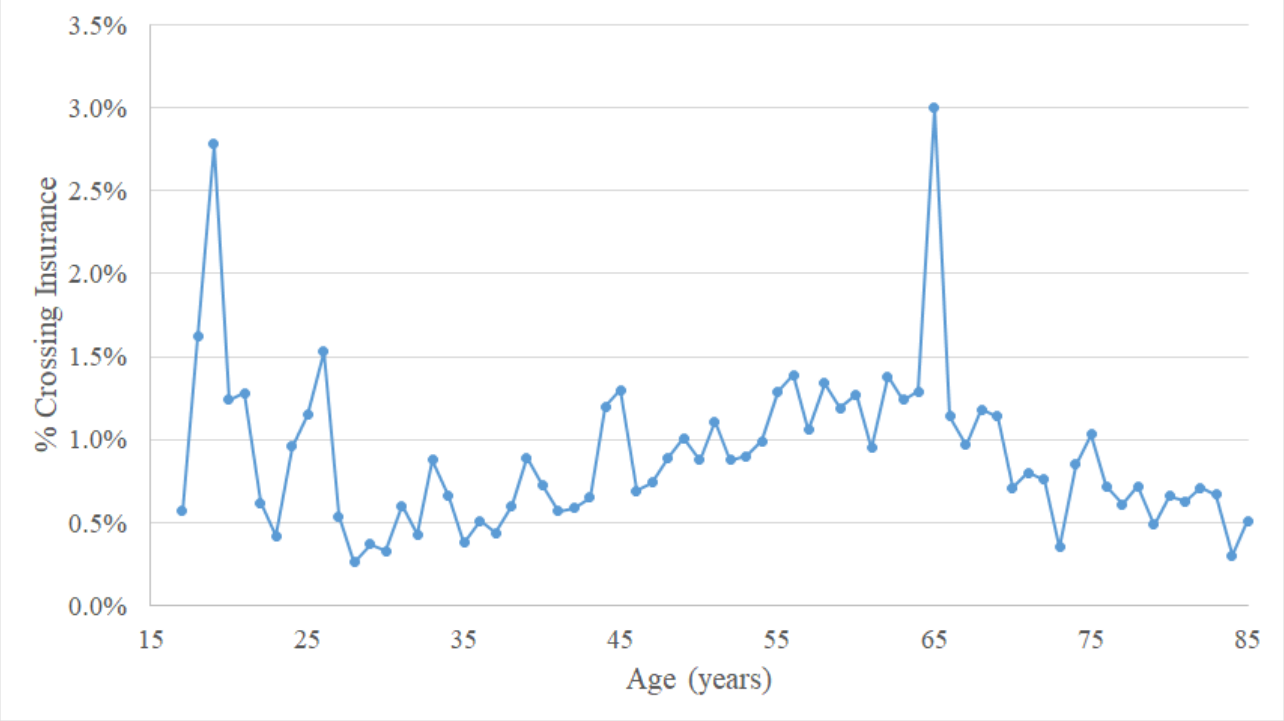


Age

Figure 3 highlights that crossover alerts occurred more often at age 19 (7/252, 2.8%; *P*=.004), 5.37 times more likely than patients at ages other than 19, 26, and 65 (95% CI 2.50-11.47; *P*<.001), and significantly more often at age 65 (193/6431, 3.0%; *P*<.001), 2.95 times more likely (95% CI 2.53-3.45; *P*<.001). The rate of crossover alerts was directionally higher in bivariate analysis (12/783, 1.5%) at age 26, but not at the level of statistical significance (*P*=.12). However, multivariate

results found that patients aged 26 years were 2.38 times more likely to have crossover alerts (95% CI 1.26-4.47; *P*=.007). Patients younger than 17 years of age seldom experienced serious drug-drug interaction alerts (720/242,646 or 0.3% of the total alerts).

Figure 3. Percentage of alerts crossing insurance identifier by age (September 1, 2016 to August 31, 2019).



Patient Movement Between Pharmacy Benefit Managers

We found 373,929 patients between 38 and 48 years of age with at least 2 claims for atorvastatin or amlodipine besylate during 2017. Of those, 69,500 (18.6%) had no claims processed by this pharmacy benefit manager during 2018, and 117,069 (31.3%) had no claims processed by this pharmacy benefit manager during 2019. We similarly found 412,101 patients who had at least 2 claims for atorvastatin or amlodipine besylate during 2019, when they were between 38 and 48 years of age. Of those, 76,222 (18.5%) had no claims processed by this pharmacy benefit manager in 2018, and 147,520 (35.8%) had no claims processed by this pharmacy benefit manager in 2017. These findings confirm our assumption that patients move regularly between pharmacy benefit managers.

Therapeutic Class

Antibiotics generated the most serious drug-drug interaction alerts overall. Macrolide antibiotics represented 22.5% (54,667/242,646) of all serious drug-drug interaction alerts, and quinolone antibiotics represented 14.5% (35,083/242,646). The second most common was opioids: opioids with nonsalicylates (eg, acetaminophen with codeine) represented 9.6% (23,323/242,646) of all serious drug-drug interaction alerts, and opioid analgesics (eg, tramadol) represented 7.9% (19,239/242,646) of all serious drug-drug interaction alerts. Third was products treating erectile dysfunction (representing 9.9%, 24,090/242,646) of all serious drug-drug interaction alerts and pulmonary arterial hypertension (representing 3.1%, 7632/242,646) of all serious drug-drug interaction alerts. Both erectile dysfunction and pulmonary arterial hypertension medications contain sildenafil and tadalafil.

The rate of crossover alerts was significantly higher (*P*<.001) among claims for erectile dysfunction medications (487/24,090, 2.0%), pulmonary arterial hypertension (171/7632, 2.2%), and vasodilators (499/20,342, 2.4%) compared to those of all other therapy classes. Multivariate analysis shows that alerts for erectile dysfunction drugs were 3.23 (95% CI 2.70-3.88) times more likely to be crossover, alerts for pulmonary arterial hypertension were 3.48 (95% CI 2.81-4.31) times more likely, and alerts for vasodilators were 3.99 (95% CI 3.35-4.77) times more likely (*P*<.001). Drug classes with lower than average crossover alert rates included macrolide antibiotics 0.8% (430/54,667; *P*<.001), quinolone antibiotics 0.76% (267/35,083; *P*<.001), opioid nonsalicylates 0.5% (126/23,323; *P*<.001), and opioid analgesics 0.5% (89/19,239; *P*<.001).

Using all therapeutic classes as the reference category, the replacement rate was higher among macrolide antibiotics (9237/54,667, 16.9%), which were 1.38 times more likely to be replaced (95% CI 1.34-1.42; *P*<.001); opioid analgesics (3541/19,239, 18.4%), which were 1.92 times more likely to be replaced (95% CI 1.80-2.05; *P*<.001); and opioid nonsalicylates (5521/23,323, 23.6%), which were 2.38 times more likely to be replaced (95% CI 2.22-2.54; *P*<.001). Fewer replacements occurred among erectile dysfunction medications (333/24,090, 1.4%), which were 0.52 times less likely to be replaced (95% CI 0.49-0.55; *P*<.001); vasodilators (233/20,342, 1.1%), which were 0.41 times less likely to be replaced (95% CI 0.38-0.44; *P*<.001); and pulmonary arterial hypertension medications (25/7632, 0.32%), which were 0.39 times less likely

to be replaced (95% CI 0.32-0.49; *P*<.001). Abandonment was infrequent for macrolide antibiotics (3615/54,667, 6.6%, *P*<.001) but common for erectile dysfunction (8946/24,090; 37.1%, *P*<.001) and pulmonary arterial hypertension (5888/7632, 77.2%, *P*<.001) medications.

Gender

Males exhibited a higher proportion of crossover alerts than females (males: 1655/132,449, 1.3%; females: 733/110,197, 0.7%; *P*<.001). However, multivariate analysis indicated results were not significant (*P*=.38)

Additional Crossover Alerts Forecasted With Complete Unique Patient Identifier and Information Exchange

The proportionality factor was (1 / 0.165) – 1= 5.06. Assuming an effective unique patient identifier and complete sharing of prescription data, crossover alerts would increase by a factor of 5.06, resulting in a crossover alert percentage of 5.0%, compared to the original 0.98% (2388/242,646). This rate would be greatest in January, when crossovers are more common. Assuming effective unique patient identifier and complete sharing of prescription data, additional crossover alerts found during January, with an observed crossover alert rate of 1.9% (414/21,801), would increase to 9.6% using the proportionality factor.

Total Estimated Annual Serious Drug-Drug Interaction Alerts Undiscovered

Using the projected crossover alert percentage of 5.0%, our results indicate that, annually, 22,730 serious drug-drug interaction alerts are undetected by the pharmacy benefit manager.

Total Estimated Annual Serious Drug-Drug Interaction Alerts That May Result in a Contraindicated Dispensing

We estimate that of the 22,730 undetected serious drug-drug interaction alerts, 6023 (26.5%) would have been replaced or abandoned had they been detected. We therefore estimate that undetected serious drug-drug interaction alerts may contribute to up to 6023 annual contraindicated dispensings because the pharmacy benefit manager does not alert the pharmacy.

Discussion

Principal Findings

A significant minority of patients moves annually within and among pharmacy benefit managers, increasing the risk for undetected serious drug-drug interaction alerts due to lack of interoperability and consistent identification. Our analysis highlights several important factors associated with these transitions. Understanding these factors both highlights the need for improved interoperability and can inform future interoperability improvements to mitigate clinical risk.

Most prescriptions are dispensed for a maximum supply of 90 days, thus precipitating claims more than 3 months old are unlikely to trigger a serious drug-drug interaction alert. Therefore, the peak in crossover alerts observed January through March may be explained by patient transition to a new, nonintegrated payor, often at the beginning of each calendar year, fragmenting prescription history. The new pharmacy benefit manager accumulates new pharmacy claims as February progresses into March. These new claims increasingly trigger their own serious drug-drug interaction alerts, while crossover alerts requiring an integrated prescription history decrease.

In addition to yearly fluctuations, health insurance transitions are heightened at specific ages. Crossover alerts increase at ages 19 and 26, resulting from transitions from a family plan, and at age 65 from transitions into Medicare.

We hypothesize that lower-income Medicaid beneficiaries, compared to those of Medicare Part D and employer plan beneficiaries, experience fewer crossover alerts for several reasons. When patients transition to Medicaid, they often have no immediate prior coverage and no associated prescription history. Patients with permanent disabilities having lifelong Medicaid eligibility are also unlikely to switch plans. Medicaid-eligible patients can apply throughout the year, and we did not observe seasonal variation in crossover alert rates for Medicaid beneficiaries. In contrast, Exchange Plan patients must choose their insurer at year-end, and they experience more crossover alerts, particularly in January.

The observed increase in crossover alerts associated with sildenafil and tadalafil may result from noncoverage. Prescription plans often deny benefits for erectile dysfunction treatment. Consequently, patients often seek alternative coverage for these medications, which results in prescription history recorded under a different beneficiary identifier. Vasodilators, which interact with sildenafil and tadalafil, produce more crossover alerts. Noncoverage of erectile dysfunction may also explain the higher rate of crossover alerts for males (24,018/24,090, 99.7% of claims for erectile dysfunction products and 5938/7632, 77.8% of pulmonary arterial hypertension products are dispensed to reported males).

We found a statistical but not clinically meaningful difference in pharmacy response to crossover versus their response to noncrossover alerts. The pharmacy benefit manager studied does not disclose to pharmacies whether a serious drug-drug interaction alert is a crossover alert. The alerted pharmacist learns that the patient has a potential conflict, not how the conflict was identified.

While we observed differences between the rates of crossover, replacement, override, and abandonment among the pharmacy chains studied, these differences were not meaningful and did not impact our conclusions. Similarly, differences, though small, were noticed in average days' supply for crossovers but did not impact our conclusions. We hypothesize that the findings of differences in mean days' supply between replacements and overrides is related to the dispensed drug. Antibiotics, as opposed to opioids and erectile dysfunction drugs, are more often replaced and more often dispensed for an acute treatment period.

Our results suggest that improved identification and medication history exchange could help pharmacy benefit managers identify up to 5.0% additional serious drug-drug interactions, and in January, up to 9.7% additional serious drug-drug interactions.

Limitations

It is possible that many serious drug-drug interaction alerts identified by pharmacy benefit managers using a unique patient identifier may also be detected by the dispensing pharmacy. We lack data to determine the proportion of serious drug-drug interaction alerts triggered by both the pharmacy benefit manager and the pharmacy, as well as the proportion identified solely by the pharmacy benefit manager. Nevertheless, it is clear that using a unique patient identifier enhances the pharmacy benefit manager's ability to identify serious drug-drug interaction alerts. That being said, we do not directly link unique patient identifier usage to improved health outcomes.

While a unique patient identifier appears to be helpful, differing deployments of unique patient identifier may have varying benefits or introduce novel problems. A unique patient identifier with false positive matches may lead to false positive serious drug-drug interaction alerts, and errors in transcribing a unique patient identifier may lead to misidentification. Furthermore, this study does not address the influence of e-prescribing, which may improve the prescriber's awareness of the patient's prescription history and thereby reduce the risk of unmanaged serious drug-drug interactions.

Many factors contribute to the feasibility of various strategies for improving identity. Chief among them is accuracy of the matching process, and the corresponding improvement in clinical outcomes. While this study evaluated the clinical outcomes that could be realized through improved identification, we do not address the issues of privacy and security, which we acknowledge have posed significant barriers to deployment of a national unique patient identifier.

Our results may not be generalizable to other health care contexts. Other providers who receive patient data in different ways will face different challenges. The impact of improved identification on clinical outcomes depends on many factors including workflow, data sources, and data quality. Thus, it is likely that providers in other roles who adopt comprehensive patient identification strategies will achieve different degrees of improvement. However, our results suggest that improved identification can improve outcomes, in this case detection of serious drug-drug interaction alerts. Estimates of impact will require experimental verification and further analysis in additional settings.

Conclusion

Because the US lacks both a comprehensive identification strategy and ubiquitous health information exchange, our results indicate that up to 6023 contraindicated codispensings may go undetected each year among insured patients. Although progress is being made in US health care systems toward more comprehensive interoperability, fragmented information silos remain the status quo. When patients transition to a new insurer or pharmacy benefit manager, their identity and historical prescription data do not seamlessly follow. Subsequently, pharmacy benefit managers may lack both identifying information and historical data.

A prospective drug utilization review process that does not rely upon a comprehensive patient identity strategy is likely to miss serious drug-drug interaction alerts. A pharmacy benefits manager with a significant market share of the US population that uses only an insurance identifier to aggregate patient records for prospective drug utilization review is likely to miss 1% or more of serious drug-drug interaction alerts, even when using patient information that they already possess but have not linked. The risk of missed serious drug-drug interaction alerts is greater when patients commonly move between benefits: each year in the month of January, and at the ages of 19, 26, and 65.

Additional alerts detected solely through the adoption of a unique patient identifier (ie, without interoperable data sharing) are likely to increase in direct proportion to the size of the population a pharmacy benefit manager system serves. A pharmacy benefit manager that serves more than 15% of the US population and begins using a unique patient identifier in prospective drug utilization review (without any new data transfers) is likely to find an additional 1% serious drug-drug interaction alerts among its patients. A pharmacy benefit manager with smaller market share is likely to identify fewer additional serious drug-drug interaction alerts.

Increased prospective drug utilization review alerting rates beyond those achieved with improved identification can likely be realized with routine information sharing between providers. If all pharmacy benefit managers comprehensively exchanged information for the purpose of prospective drug utilization reviews, it is likely that pharmacy benefit managers would identify 5% more serious drug-drug interaction alerts.

However, serious drug-drug interaction alerts in the future will not necessarily be discovered at the pharmacy counter. As electronic health record systems capture more pharmacy claims data, physician office visits should benefit from more complete patient medication history. This additional data may enable electronic health records to identify more serious drug-drug interaction risks before transmitting a prescription to the pharmacy.

In order to minimize serious drug-drug interactions, we must ensure that comprehensive medication history data is available to prospective drug utilization review. Pharmacy benefit managers that have not implemented a comprehensive patient identity management strategy for prospective drug utilization review should consider doing so. In similar fashion, stakeholders across the health care spectrum should consider implementing comprehensive patient identity management and information exchange strategies to minimize medical errors due to incomplete and missing data.

In order to identify more serious drug-drug interaction alerts, pharmacy benefit managers must ensure that prescribing history is available at times of transition. To do so, pharmacy benefit managers should routinely share new patients' prior history, regardless of whether transitioning payors request the transfer. Other providers in the health care community should also plan to use interoperability standards in order to obtain relevant records about each patient before providing services.

In the short term, until these measures are achieved, pharmacists should be aware that automated prospective drug utilization review is likely to miss nearly 10% of serious drug-drug

interaction alerts in January. They should take particular care during January, ensuring awareness of patients with new coverage using potentially conflicting medications.

While issues of privacy and security remain to be addressed, our data shows that consistent identification can help identify additional serious drug-drug interactions. Given the volume of opportunities to improve patient care, the health care system should choose the most accurate identification strategy possible. We hope that others will conduct similar studies in other areas of the health care ecosystem to forecast benefits from patient identification and patient-record sharing.

Conflicts of Interest

None declared.

References

1. Adams K, Howe J, Fong A, Puthumana J, Kellogg K, Gaunt M, et al. An Analysis of Patient Safety Incident Reports Associated with Electronic Health Record Interoperability. Appl Clin Inform 2017 Dec 21;08(02):593-602. [doi: 10.4338/aci-2017-01-ra-0014]
2. Culbertson A, Goel S, Madden M, Safaeinili N, Jackson K, Carton T, et al. The Building Blocks of Interoperability. A Multisite Analysis of Patient Demographic Attributes Available for Matching. Appl Clin Inform 2017 Apr 05;8(2):322-336 [FREE Full text] [doi: 10.4338/ACI-2016-11-RA-0196] [Medline: 28378025]
3. Reisman M. EHRs: The Challenge of Making Electronic Data Usable and Interoperable. P T 2017 Sep;42(9):572-575 [FREE Full text] [Medline: 28890644]
4. Louis CJ, Clark JR, Hillemeier MM, Camacho F, Yao N, Anderson RT. The Effects of Hospital Characteristics on Delays in Breast Cancer Diagnosis in Appalachian Communities: A Population-Based Study. J Rural Health 2018 Feb;34 Suppl 1:s91-s103 [FREE Full text] [doi: 10.1111/jrh.12226] [Medline: 28102909]
5. Challenges and Strategies for Accurately Matching Patients to their Health Data Internet. Bipartisan Policy Center. 2012. URL: https://bipartisanpolicy.org/wp-content/uploads/2019/03/BPC-HIT-Issue-Brief-on-Patient-Matching.pdf [accessed 2020-02-10]
6. CMS Interoperability and Patient Access final rule. Centers for Medicare & Medicaid Services. 2020. URL: https://www.cms.gov/Regulations-and-Guidance/Guidance/Interoperability/index [accessed 2020-09-20]
7. Culbertson A, Goel S, Madden M, Safaeinili N, Jackson K, Carton T, et al. The Building Blocks of Interoperability. A Multisite Analysis of Patient Demographic Attributes Available for Matching. Appl Clin Inform 2017 Apr 05;8(2):322-336 [FREE Full text] [doi: 10.4338/ACI-2016-11-RA-0196] [Medline: 28378025]
8. Labor, Health and Human Services, Education, Defense, State, Foreign Operations, and Energy and Water Development Appropriations Act, 2020. H.R. 2740, 116th Congress. Washington, DC: United States Congress; 2019.
9. Labor, Health and Human Services, Education, Defense, State, Foreign Operations, and Energy and Water Development Appropriations Act, 2020. H.Amdt. 296 to H.R. 2740, 116th Congress. Washington, DC: United States Congress; 2019.
10. FY2020 Consolidated Appropriations (Labor-HHS-ED, Defense, State-Foreign Operations, Energy and Water) H.R. 2740, 116th Congress. Washington, DC: United States Congress; 2019.
11. FY2020 Further Consolidated Appropriations Act (LHHSED, AG, Energy Water, Interior, Leg. Branch, MCVA, State-For. Ops, T-HUD), H.R. 1865, 116th Congress. Washington, DC: United States Congress; 2019.
12. United States Congress. Further Consolidated Appropriations Act, 2020. H.R. 1865/Public Law 116-94, 116th Congress. Washington, DC: United States Congress; 2019.
13. Wiggins BS, Saseen JJ, Page RL, Reed BN, Sneed K, Kostis JB, et al. Recommendations for Management of Clinically Significant Drug-Drug Interactions With Statins and Select Agents Used in Patients With Cardiovascular Disease: A Scientific Statement From the American Heart Association. Circulation 2016 Nov 22;134(21):e468-e495. [doi: 10.1161/cir.0000000000000456]
14. Shapiro L, Shear N. Drug-drug interactions: how scared should we be? CMAJ 1999;161:1266-1267. [doi: 10.1016/b978-0-323-61211-1.00066-8]
15. Jankel CA, Fitterman LK. Epidemiology of Drug-Drug Interactions as a Cause of Hospital Admissions. Drug Safety 1993;9(1):51-59. [doi: 10.2165/00002018-199309010-00005]
16. Schneitman-McIntire O, Farnen T, Gordon N, Chan J, Toy W. Medication misadventures resulting in emergency department visits at and HMO medical center. Am J Health Syst Pharm 1996 Jun 15;53(12):1416-1422. [doi: 10.1093/ajhp/53.12.1416] [Medline: 8781687]
17. Hamilton RA, Briceland LL, Andritz MH. Frequency of hospitalization after exposure to known drug-drug interactions in a Medicaid population. Pharmacotherapy 1998;18(5):1112-1120. [Medline: 9758323]
18. Rucker NL. New federal DUR requirements for pharmacists in 1993. Am Pharm 1992 Sep;NS32(9):44-6, 58. [doi: 10.1016/s0160-3450(15)31003-5] [Medline: 1442554]

19. Edrees H, Amato M, Wong A, Seger D, Bates D. High-priority drug-drug interaction clinical decision support overrides in a newly implemented commercial computerized provider order-entry system: Override appropriateness and adverse drug events. J Am Med Inform Assoc 2020 Jun 01;27(6):893-900. [doi: 10.1093/jamia/ocaa034] [Medline: 32337561]
20. Vasquez J, Lohr G. Pharmacy benefit managers, explained. www.advisory.com.: www.advisory.com; 2019. URL: https://www.advisory.com/daily-briefing/2019/11/13/pbms [accessed 2020-09-20]
21. Bailey JE, Wan JY, Mabry LM, Landy SH, Pope RA, Waters TM, et al. Does health information exchange reduce unnecessary neuroimaging and improve quality of headache care in the emergency department? J Gen Intern Med 2013 Feb;28(2):176-183 [FREE Full text] [doi: 10.1007/s11606-012-2092-7] [Medline: 22648609]
22. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood) 2005;Suppl Web Exclusives:W5-10. [doi: 10.1377/hlthaff.w5.10] [Medline: 15659453]
23. Dunn EJ, Moga PJ. Patient misidentification in laboratory medicine: a qualitative analysis of 227 root cause analysis reports in the Veterans Health Administration. Arch Pathol Lab Med 2010 Feb;134(2):244-255 [FREE Full text] [doi: 10.1043/1543-2165-134.2.244] [Medline: 20121614]
24. Schulmeister L. Patient misidentification in oncology care. Clin J Oncol Nurs 2008 Jun;12(3):495-498 [FREE Full text] [doi: 10.1188/08.CJON.495-498] [Medline: 18515248]
25. Lippi G, Mattiuzzi C, Bovo C, Favaloro EJ. Managing the patient identification crisis in healthcare and laboratory medicine. Clin Biochem 2017 Jul;50(10-11):562-567. [doi: 10.1016/j.clinbiochem.2017.02.004] [Medline: 28179154]
26. Gray JE, Suresh G, Ursprung R, Edwards WH, Nickerson J, Shiono PH, et al. Patient misidentification in the neonatal intensive care unit: quantification of risk. Pediatrics 2006 Jan;117(1):e43-e47. [doi: 10.1542/peds.2005-0291] [Medline: 16396847]
27. Pfeifer E, Lozovatsky M, Abraham J, Kannampallil T. Effect of an Alternative Newborn Naming Strategy on Wrong-Patient Errors: A Quasi-Experimental Study. Appl Clin Inform 2020 Mar;11(2):235-241. [doi: 10.1055/s-0040-1705175] [Medline: 32236916]
28. Patient Identification Errors. ECRI Institute. 2016. URL: https://www.ecri.org/Resources/HIT/Patient ID/Patient_Identification_Evidence_Based_Literature_final.pdf [accessed 2020-09-20]
29. Moscovitch B, Halamka J, Grannis S. Better patient identification could help fight the coronavirus. NPJ Digit Med 2020;3:83 [FREE Full text] [doi: 10.1038/s41746-020-0289-4] [Medline: 32529044]
30. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. Proc AMIA Symp 2002:305-309 [FREE Full text] [Medline: 12463836]
31. Berchick ER, Barnett JC, Upton RD. Health Insurance Coverage in the United States: 2018. United States Census Bureau. 2019. URL: https://www.census.gov/library/publications/2019/demo/p60-267.html [accessed 2020-09-20]
32. Actuarial Life Table. Social Security Administration. 2020. URL: https://www.ssa.gov/oact/STATS/table4c6.html [accessed 2020-09-20]

Edited by G Eysenbach; submitted 17.08.20; peer-reviewed by B Protus, S Wilson; comments to author 15.09.20; revised version received 14.10.20; accepted 30.10.20; published 20.11.20.

Please cite as:
Sragow HM, Bidell E, Mager D, Grannis S
Universal Patient Identifier and Interoperability for Detection of Serious Drug Interactions: Retrospective Study
JMIR Med Inform 2020;8(11):e23353
URL: http://medinform.jmir.org/2020/11/e23353/
doi:10.2196/23353
PMID:33216009

Original Paper

Explainable Artificial Intelligence Recommendation System by Leveraging the Semantics of Adverse Childhood Experiences: Proof-of-Concept Prototype Development

Nariman Ammar[1], PhD; Arash Shaban-Nejad[1], MSc, MPH, PhD

University of Tennessee Health Science Center - Oak Ridge National Laboratory, Center for Biomedical Informatics, Department of Pediatrics, College of Medicine, Memphis, TN, United States

Corresponding Author:
Arash Shaban-Nejad, MSc, MPH, PhD
University of Tennessee Health Science Center - Oak Ridge National Laboratory, Center for Biomedical Informatics
Department of Pediatrics, College of Medicine
Memphis, TN
United States
Phone: 1 901 287 583
Email: ashabann@uthsc.edu

Abstract

Background: The study of adverse childhood experiences and their consequences has emerged over the past 20 years. Although the conclusions from these studies are available, the same is not true of the data. Accordingly, it is a complex problem to build a training set and develop machine-learning models from these studies. Classic machine learning and artificial intelligence techniques cannot provide a full scientific understanding of the inner workings of the underlying models. This raises credibility issues due to the lack of transparency and generalizability. Explainable artificial intelligence is an emerging approach for promoting credibility, accountability, and trust in mission-critical areas such as medicine by combining machine-learning approaches with explanatory techniques that explicitly show what the decision criteria are and why (or how) they have been established. Hence, thinking about how machine learning could benefit from knowledge graphs that combine “common sense” knowledge as well as semantic reasoning and causality models is a potential solution to this problem.

Objective: In this study, we aimed to leverage explainable artificial intelligence, and propose a proof-of-concept prototype for a knowledge-driven evidence-based recommendation system to improve mental health surveillance.

Methods: We used concepts from an ontology that we have developed to build and train a question-answering agent using the Google DialogFlow engine. In addition to the question-answering agent, the initial prototype includes knowledge graph generation and recommendation components that leverage third-party graph technology.

Results: To showcase the framework functionalities, we here present a prototype design and demonstrate the main features through four use case scenarios motivated by an initiative currently implemented at a children’s hospital in Memphis, Tennessee. Ongoing development of the prototype requires implementing an optimization algorithm of the recommendations, incorporating a privacy layer through a personal health library, and conducting a clinical trial to assess both usability and usefulness of the implementation.

Conclusions: This semantic-driven explainable artificial intelligence prototype can enhance health care practitioners’ ability to provide explanations for the decisions they make.

(JMIR Med Inform 2020;8(11):e18752) doi:10.2196/18752

KEYWORDS

mental health surveillance; semantic web; knowledge-based recommendation; digital assistant; explainable artificial intelligence; adverse childhood experiences

Introduction

Background

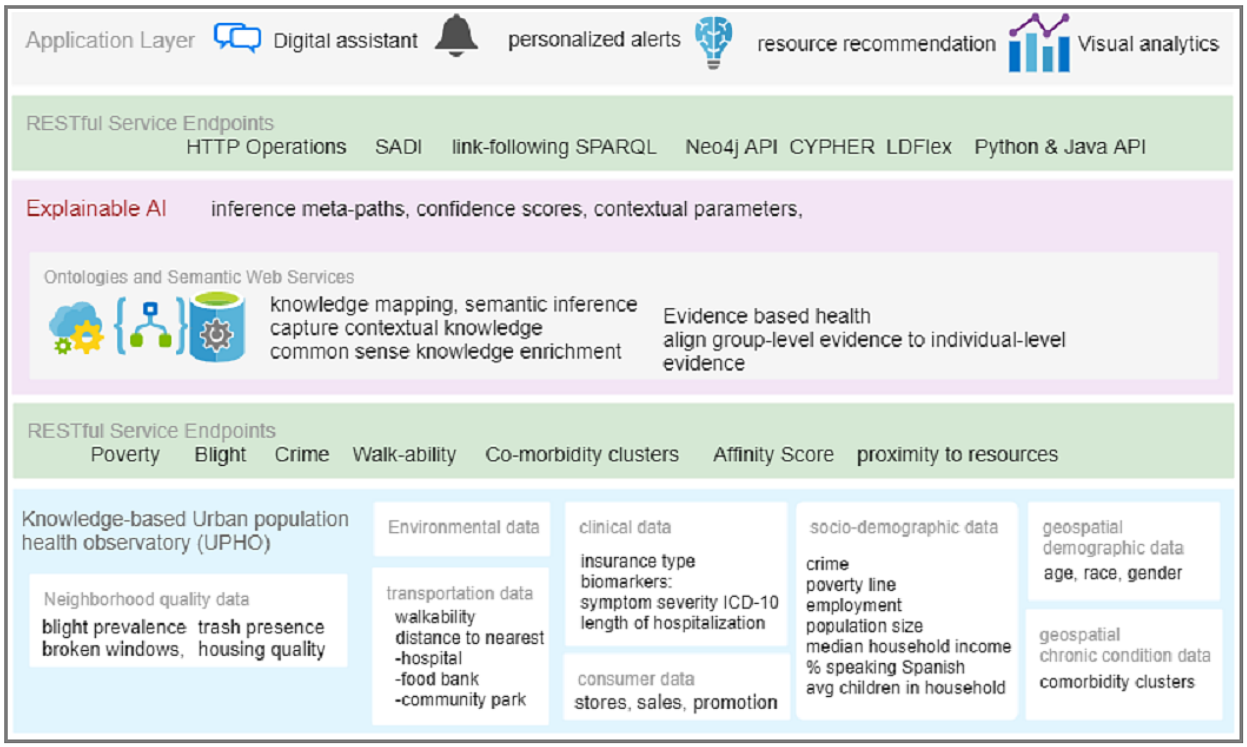
The concept of adverse childhood experiences (ACEs) has been recognized for quite some time but was first formally studied in the CDC-Kaiser landmark study [1], which uncovered the strong connection between ACEs and the development of risk factors for different negative health outcomes that threaten the well-being of populations throughout their life course. Social determinants of health (SDoH) are measurable indicators of social conditions in which a patient is embedded. Individuals who experience a more negative burden of these factors within their neighborhood are at higher risk of negative health outcomes [2-4]. There is an entire body of research focused on studying the links between ACEs and SDoH and health outcomes, but few intelligent tools are available to assist in the real-time screening of patients and to assess the connection between ACEs and SDoH, which could help to guide patients and families to available resources (eg, health care providers, government, and nongovernment agencies). Other recent works have focused on developing question-answering (QA) systems for training nursing practitioners on how to answer patient inquiries [5].

Recommendation systems and digital assistants often require machine learning (ML), artificial intelligence (AI), and natural language processing capabilities to effectively connect and harvest the vast amounts of generated data. They also need to store, retrieve, and learn from past interactions and experiences with users. Traditionally, recommendation systems have relied on classic ML techniques that often cannot provide a full scientific understanding of the inner workings of the underlying models. For AI to mimic human intelligence, it needs to incorporate one of the most important classes of information people use to make predictions and decisions: context. Although a standard AI algorithm can learn useful rules from a training set, it also tends to learn other unnecessary, nongeneralizable rules, which may lead to a lack of consistency, transparency, and generalizability. Explainable AI is an emerging approach for promoting credibility, accountability, and trust in mission-critical areas such as medicine by combining ML techniques with explanatory techniques that explicitly show why a recommendation is made. One way to achieve this is by considering formal ontologies as an integral part of the learning process, providing the necessary contextual knowledge about a phenomenon. AI and ML become more trustworthy when underpinned by contextual information provided by ontological platforms. Ontologies can be represented through a graph structure. When computations are performed over graph-structured data, they can make generalizations related to structure rather than data since they support relational reasoning and combinatorial generalization [6]. Moreover, when AI apps are based on contextually aware and dynamic backends, they become easier to train with minimal maintenance. A knowledge graph [7] can serve as a dynamic backend that stores data in a certain domain as entities and relationships using a graph model, which abides by an ontology. Several studies have incorporated graph technologies into ML models applied to biomedical informatics problems [8-11].

We here propose the Semantic Platform for Adverse Childhood Experiences Surveillance (SPACES), an intelligent recommendation system that employs ML techniques to help in screening patients and allocating or discovering relevant resources. The novelty in the approach lies in its ability to use the contextual knowledge collected about the user, and infer new knowledge to support subsequent QA and resource allocations during intake assessment sessions in (near) real time. Moreover, our proposed system intends to build rapport with patients by generating personalized questions during interviews while minimizing the amount of information that needs to be collected directly from the patient.

In our previous work [12-14], we developed the Adverse Childhood Experiences Ontology (ACESO) that captures knowledge on ACEs, SDoH, health outcomes, and interventions. Both expressive and light versions of the ontology can be freely downloaded via BioPortal [15]. The ontology defines concept and property hierarchies, and encodes causal epidemiological knowledge as axioms. We also developed a repository termed Urban Population Health Observatory (UPHO), which provides metrics based on several socioeconomic and environmental data at the neighborhood level that can be linked to health outcomes. Figure 1 demonstrates the multiple layers, from data to the app, that are required for building a knowledge-based explainable and interpretable model.

Figure 1. Multilayer representation of a knowledge-based explainable/interpretable model.



ACESO

There are two types of knowledge captured in the ontology: (1) domain concepts, in which the ontology encodes concepts about risk factors, including ACEs (eg, abuse), SDoH (eg, housing condition), health outcomes such as chronic diseases (eg, asthma) and stress, and interventions; and (2) semantic inference, in which some of the knowledge in the ontology is encoded in the form of axioms. The axioms express knowledge related to (1) inclusions that define how two concepts are related; (2) equivalence relationships; and (3) causal knowledge, which are statements in the form of assertions of links between risk factors and health outcomes (eg, "obesity is a risk factor of diabetes" and "stress and exposure to toxins are risk factors for asthma").

Knowledge-Based UPHO

The UPHO is a knowledge-based repository that can be used to represent and infer neighborhood-level indicators (eg, blight prevalence, poverty, education, proximity to clinics, proximity to public transportation) that may lead to negative health outcomes (eg, asthma, diabetes, stress, obesity). Moreover, the UPHO provides an analytics layer for calculating several metrics (eg, affinity scores in chronic conditions as a measure for comorbidity [16]) from analyzing neighborhood data.

Objectives

We aimed to develop a knowledge-driven evidence-based recommendation system and a digital assistant to facilitate mental health surveillance. We first present our methodology along with the general architecture of the proposed recommendation system. We then demonstrate the feasibility and usability of our approach through multiple use case scenarios and offer recommendations for further development.

Methods

Platform Design

The idea behind the SPACES platform is to monitor the causes of ACEs and SDoH, and their impacts on health. This platform is based on the ACESO to provide the contextual knowledge needed to facilitate intelligent exploratory and explanatory analysis. Through this framework, decision makers can (1) identify risk factors, (2) integrate and validate ACEs and SDoH exposure at individual and population levels, (3) and detect high-risk groups. The idea for implementing the recommendation system for surveillance of ACEs was motivated by a study conducted under the Family Resilience Initiative (FRI) at Le Bonheur Children's Hospital (Memphis, Tennessee) that serves families with children during regular child visits in the clinic. Our use case scenarios to demonstrate the main components and features that constitute the SPACES framework were inspired by client examples and typical issues (Textbox 1), and the follow-up activities (Textbox 2) reported in the FRI reports.

Textbox 1. Typical adverse childhood experiences (ACEs)- and social determinants of health (SDoH)-related risk factors that arise in Family Resilience Initiative reports.

ACEs

- Child behavioral issues
- Child developmental health

SDoH

- Housing
- Food insecurity
- Transportation
- Education
- Legal/benefits

Textbox 2. Typical follow-up activities in Family Resilience Initiative reports.

Well-being check-in

Following up on a referral

Renewal inquiry

Client assistance

Contact resources on behalf of a client

Sharing information about future training

Confirming appointment (psychological, clinical, education, legal)

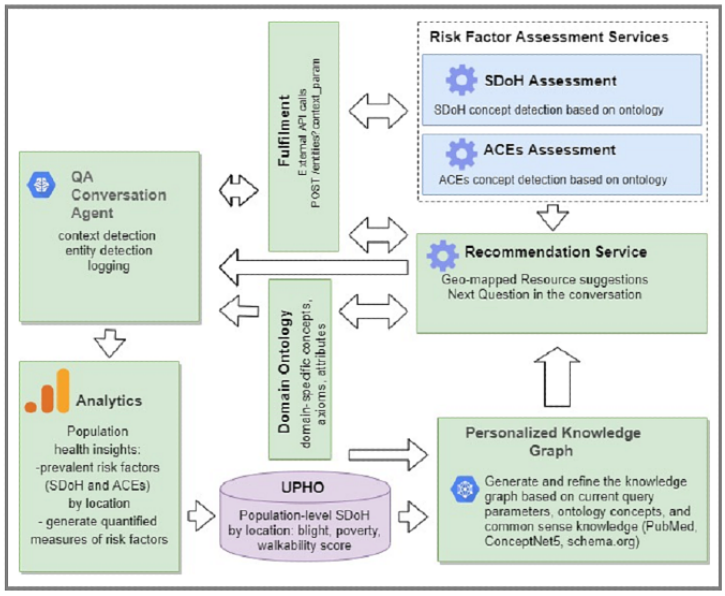
Arranging transportation

Scheduling appointment (psychological, clinical, education, legal)

System Architecture

The main components of the SPACES framework are illustrated in Figure 2. A detailed explanation of each component is provided below.

Figure 2. Architecture of the Semantic Platform for Adverse Childhood Experiences (SPACEs) that reflects the main components of the system. The blue components are the services that we implement for the mental health domain, and the green components reflect general components that can be generalized to any other domain. QA: question-answering; ACEs: adverse childhood experiences; SDoH: social determinants of health.



QA Conversation Agent

To implement the conversation agent, we used the Google DialogFlow framework [17] and defined the following constructs.

Intents

An intent represents the purpose of a user's input. We start by defining a set of intents and supplying those with training phrases. This trains the conversation agent on detecting an intent based on values (eg, mold) that represent entity types (eg, @housing_circumstance) tagged in the text. For our task, we define a generic FRI_Assessment intent that is detected when the user enters a text similar to a training phrase. This intent has the following child intents corresponding to different risk factors (eg, housing, food, transportation) and follows up associated activities.

The SDoH_surveillance intent is detected whenever the text has phrases related to entity types that match SDoH-related concepts in the ontology. For instance, the entity type housing_circumstances is detected whenever the entity values "mold," "lead-based paint," "inadequate heating," and others are tagged in the text. Since all of these concepts are SDoH-related risk factors, the SDoH_Surveillance follow-up intent is also detected.

The ACEs_surveillance intent is detected whenever the text includes phrases related to entity types that match ACEs-related concepts in the ontology.

The FRI_followup_activity intent is triggered whenever an action is detected in the text (eg, schedule an appointment).

Entity Types

Entity types are ontological concepts that dictate how data are extracted from the user's raw text. For instance, the entity type housing_circumstance is detected whenever the entity values "mold," "lead-based paint," "inadequate heating," and others are tagged in the text. We load entity types from the ontology into the agent, and then enhance the agent with a minimal set of training phrases to enable it to tag entity types that appear in the phrases. Figure 3 shows a sample training phrase tagged with entity types.

Figure 3. A sample training phrase and detected contextual parameters with entity types and values.

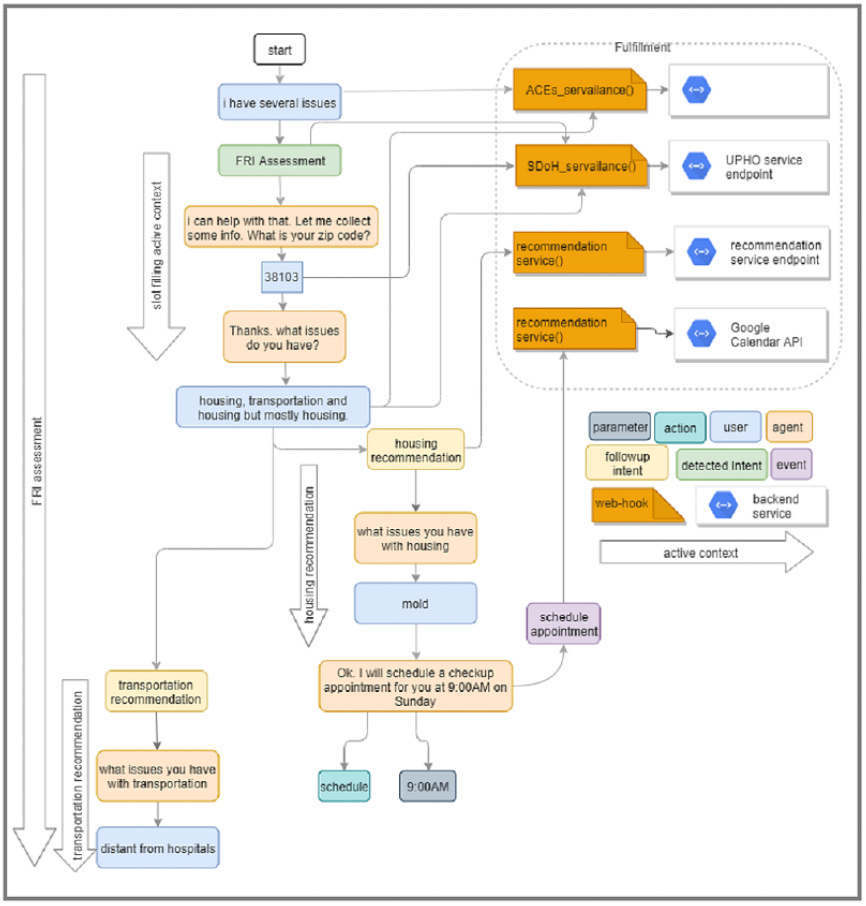
Contextual Parameters

Parameters are structured data extracted from raw text and have types that correspond to the entity types defined in the ontology. When an intent is matched at runtime, the agent extracts those parameters from entity type values (eg, 38103=zip code; night terror=symptom) that appear in the user expression (Figure 3). The agent then uses those parameters to perform logic and generate responses, or they can be exchanged between different contexts to control the conversation flow.

Contexts

To keep track of the conversation flow, the agent maintains all active contexts (see connection arrows in Figure 4) on a stack to make sure they remain active throughout the conversation. At each point in the conversation, either a new intent or a follow-up child intent is detected due to a user's input or an event and fulfillment configurations within those intents. There are different types of contexts: (a) detected intent contexts, which are triggered by training phrases; (b) follow-up intent contexts, which are triggered by their parent intent or whenever the parent intent is triggered for fulfillment; (c) slot-filling context, which becomes active when the user does not provide values for preconfigured mandatory parameters; and (d) input/output contexts, in which intents can be configured with input and output contexts. A parent intent is by default an input context for its child intent and a follow-up intent is an output context for its parent.

Figure 4. Sample conversation flow to demonstrate the different constructs used to define the question-answering (QA) agent, including intents, contexts, events, fulfillments, and webhooks. ACEs: adverse childhood experiences; UPHO: Urban Population Health Observatory; FRI: Family Resilience Initiative; SDoH: social determinants of health.



Fulfillments and Webhooks

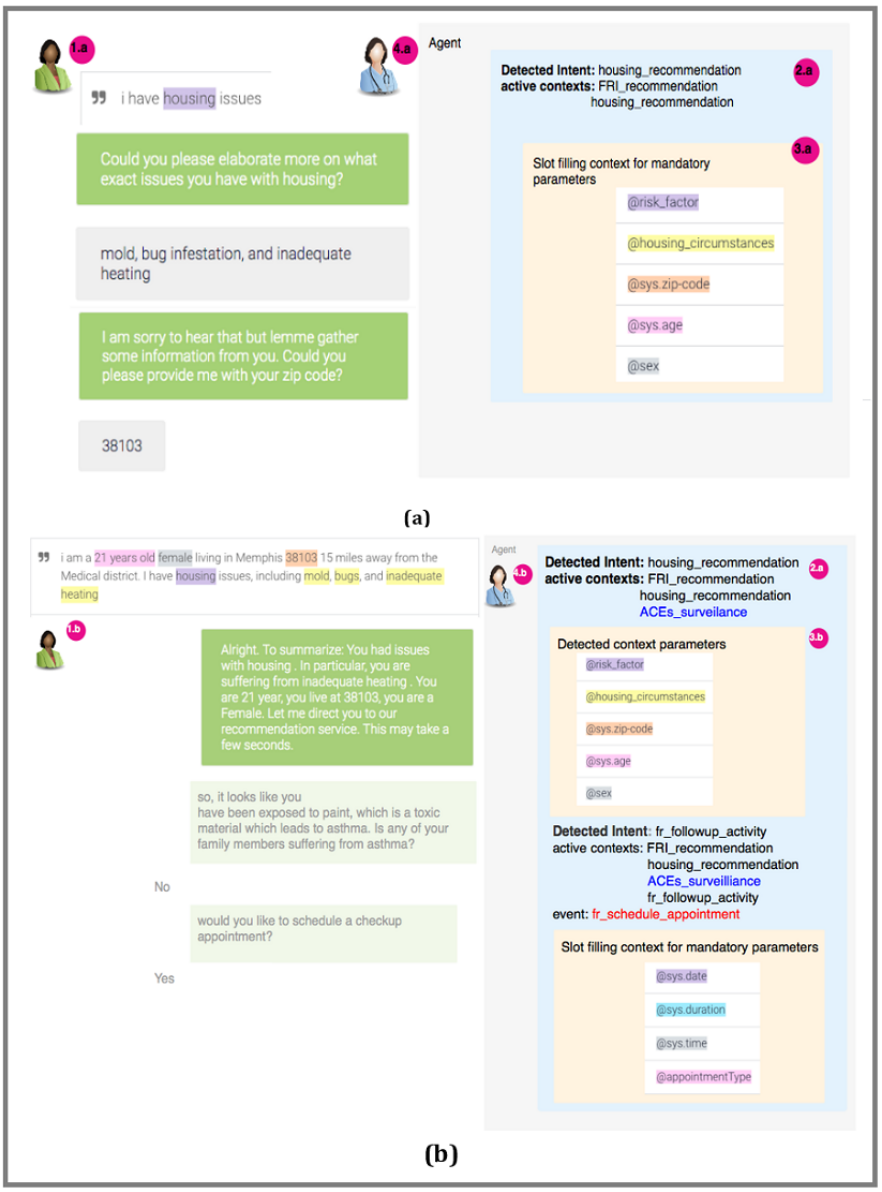
Fulfillments and webhooks enable the agent to invoke external service endpoints and send dynamic responses based on user expressions as opposed to hard-coding those responses. Fulfillment for an intent can be enabled by setting up a webhook, which is a service endpoint that we create and host. The agent sends a webhook request message that contains information about the matched intent, action, parameters, and response defined for the intent to one of our webhook services. The webhook service performs actions as needed (eg, query the knowledge graph or invoke external application programming interfaces [APIs]). The service then sends a webhook response message to the agent, which sends it to the end user.

Events

An intent could be detected either by a phrase in a user's text or by being configured for an event. Fulfillments can be used to invoke external APIs. When the agent receives a webhook response (from a backend API) that includes an event, it immediately triggers the intent in which that event is defined.

Figure 5 illustrates a sample conversation flow that shows how intent detection and context activation occur, and how the different entities within the QA agent communicate

Figure 5. Two scenarios for conversation flow. (a) Scenario 1: client does not provide much detail and is prompted with mandatory parameters asked in a certain order. (b) Scenario 2: The client provides details for mandatory parameters. Since the user's text contains lead-based paint, the agent makes a hypothesis that a household member might need an early diagnosis for asthma based on the fact that lead-based paint is a toxicant and that exposure to toxicants may lead to asthma. The parameter values get substituted at run time and more intents get detected while active contexts remain and new contexts get added. To keep track of where the user is in the conversation, the agent keeps track of all active contexts on a stack.



The user types or says an expression, which might be either detailed (eg, "I am an African American female and I have housing issues") or vague (eg, "I have several issues"). The presence or lack of details triggers different contexts. In either case, the agent matches the user expression to the generic FRI_Assessment intent, which is configured with a fulfillment that enables the agent to send a webhook request to one of the webhook services (ie, ACEs surveillance, SDoH surveillance, recommendation service). In the case of vague user input, the agent detects a slot-filling context by prompting the user with extra questions until they have provided values for all required parameters. In the case of a detailed user input, the agent directly moves to a follow-up intent, which might as well be configured with a fulfillment. After filling values for contextual parameters, the agent sends a webhook request to the recommendation service. The service responds with a webhook response that includes either a resource or a follow-up question based on the knowledge graph. A phrase provided by a user may contain parameters (eg, zip code), actions (eg, schedule an appointment), or priority words (eg, "I am interested in furthering my education, but would prefer a job first"). If a user's input includes more than one issue, then both follow-up intents are detected, but the agent handles them one at a time either based on the order they were mentioned or by using priority words.

Since the FRI_Assessment is configured with a fulfillment, the agent proceeds as though the end user initiated the match for the FRI_followup_activity intent. Thus, instead of responding to the user for the FRI_Assessment intent match, the agent

triggers the FRI_followup_activity intent, which is configured for the event "schedule an appointment." Finally, the FRI_followup_activity intent handles the required parameters (date, duration, time, and appointment type) and fulfillment (eg, Google calendar API) as dictated by the configuration of FRI_followup_activity intent.

SDoH Surveillance Service

The SDoH surveillance service is triggered whenever an SDoH-related entity type is introduced in the user's conversation with the QA agent. To achieve this, the surveillance uses the ACESO to infer whether a detected entity type (eg, being_exposed_to_lead-based_paint) is a subtype of the more generic SDoH type as well as the causal knowledge (eg, lead-based_paint *causes* asthma). This service also invokes the UPHO service endpoint to obtain metrics (eg, blight prevalence, walkability score) based on the current user's neighborhood.

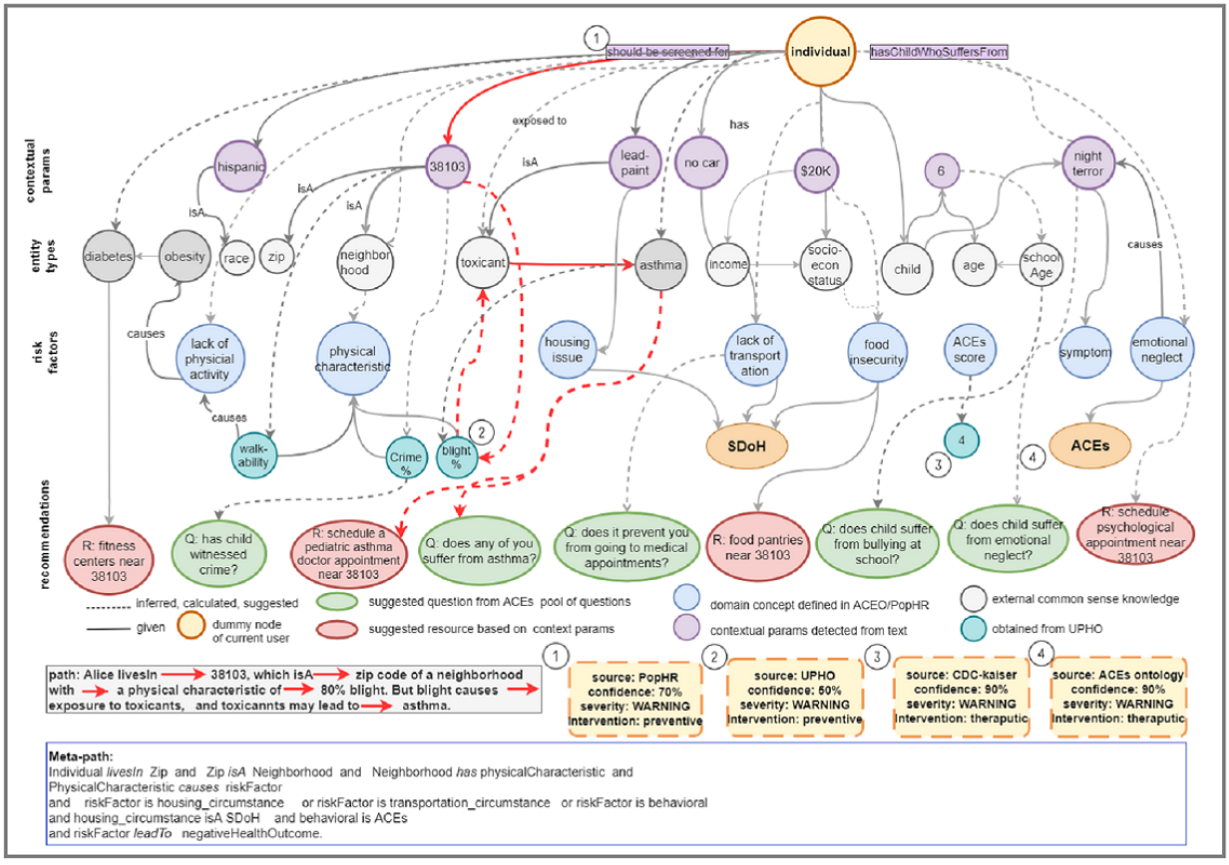
ACEs Surveillance Service

This service keeps track of all possible questions that can be asked during the ACEs assessment process [18]. It is triggered each time an ACE-related entity type is introduced in the user's conversation with the agent. It gets invoked by the recommendation service to retrieve only the questions relating to the ACEs concepts provided by the agent. The QA agent keeps track of how many ACEs questions were reported as positive, and provides those to the ACEs surveillance service, which uses the knowledge stored in the ACESO about ACEs score classification and rules for calculating them. It then sends the resulting score to the recommendation service, which seeks insights from external knowledge (eg, research publications [18]) to formulate a diagnosis based on the question: "given an ACEs score of X and symptoms S1,…Sn, what are the likely negative outcomes to screen for?"

Recommendation Service

To make real-time recommendations, this service instantly captures new resource interests, ACEs, and SDoH-associated risk factors detected in the user's current conversation and uses them to incrementally refine a personalized knowledge graph. At each stage in the conversation, the QA agent passes detected entity types and contextual parameter values to the recommendation service. Entity types help the service determine entry points on the knowledge graph, and contextual parameters help refine the queries further to obtain a more personalized version of the graph (Figure 6). Once the personalized graph is generated, the service supplies the QA agent with two types of recommendations: the next question to ask and a resource to suggest.

Figure 6. A personalized graph is generated based on contextual parameters provided by Alice (see Scenario 4 in Textbox 3) after populating ontology concepts with real-time contextual parameters supplied by the agent, and after enriching the graph with external common sense knowledge. The figure illustrates how a concrete path on the graph leads to a recommendation and the metapath that can be derived from that path.



We used Neo4j graph technology [19] for generating and enriching the graph. Before populating the graph with real-time data, we loaded ACEs, SDoH, and health outcome concepts from the ontology into a graph and persist them. We used the NeoSemantics framework [19] to import the Resource Description Framework (RDF)-based data model schemas of

the ontology as a metagraph into a Neo4j graph. The resulting property graph inherits the modeling limitations of the RDF, including the lack of support for attributes on relationships. Therefore, we enriched and fixed the raw graph after loading it in Neo4j. To populate the graph, the service starts with a *dummy node* that represents the current user (Figure 6) and then incrementally adds the following node types.

Domain concept nodes correspond to *entity types* detected by the QA agent (eg, race, age, zip code, symptom). Entity type nodes could be either (a) *risk factors* such as "housing_circumstance," "food_issue," and "household_issue," or (b) common sense knowledge such as school age and domain concepts in the given context (eg, age *is a* school age and lead-based paint *is a* toxicant).

Value nodes are populated with real-time values either (a) from *contextual parameters* obtained from the QA agent (eg, "38103," "night terror"), (b) calculated based on the ontological axioms (eg, ACEs score of 4), or (c) retrieved from the UPHO (eg, 80% blight prevalence).

Recommendation nodes are *question* nodes and *resource* nodes. Question nodes represent follow-up questions suggested for the QA agent. The answers to these questions coming from the QA agent can further refine the graph. The suggested question nodes are solicited from a pool of questions kept by the ACEs surveillance service. Resource nodes (eg, referrals, housing options, school information, doctors, clinics) are resources suggested for the QA agent to provide to the user. These resources are suggested based on the user's demographic profile (eg, the food banks within a small radius from their zip code). Resource nodes are pulled from a pool of existing resources or by invoking external backend APIs. In addition to the nodes pulled from contextual parameters, the graph shows nodes that are added based on knowledge inferred through axioms (dashed arrows in Figure 6) and concept hierarchies (solid arrows in Figure 6).

Population Health and Policymaking Analytics

The analytics component utilizes the conversation history recorded by the DialogFlow logging API, including (1) timestamps, number of interactions (within a conversation) for all user sessions, and percentage of mismatches if any; (2) a visual summary of the conversation flow with percentages for each detected intent as well as the conversational paths that users have taken when interacting with an agent; (3) popularity per intent by showing the number of sessions in which the intent was matched as well as the number of times the intent was used (total from all sessions); (4) percentage of sessions in which a user exited the conversation in the specified intent compared to the total number of sessions in which this same intent was matched; and (5) average response time to user requests.

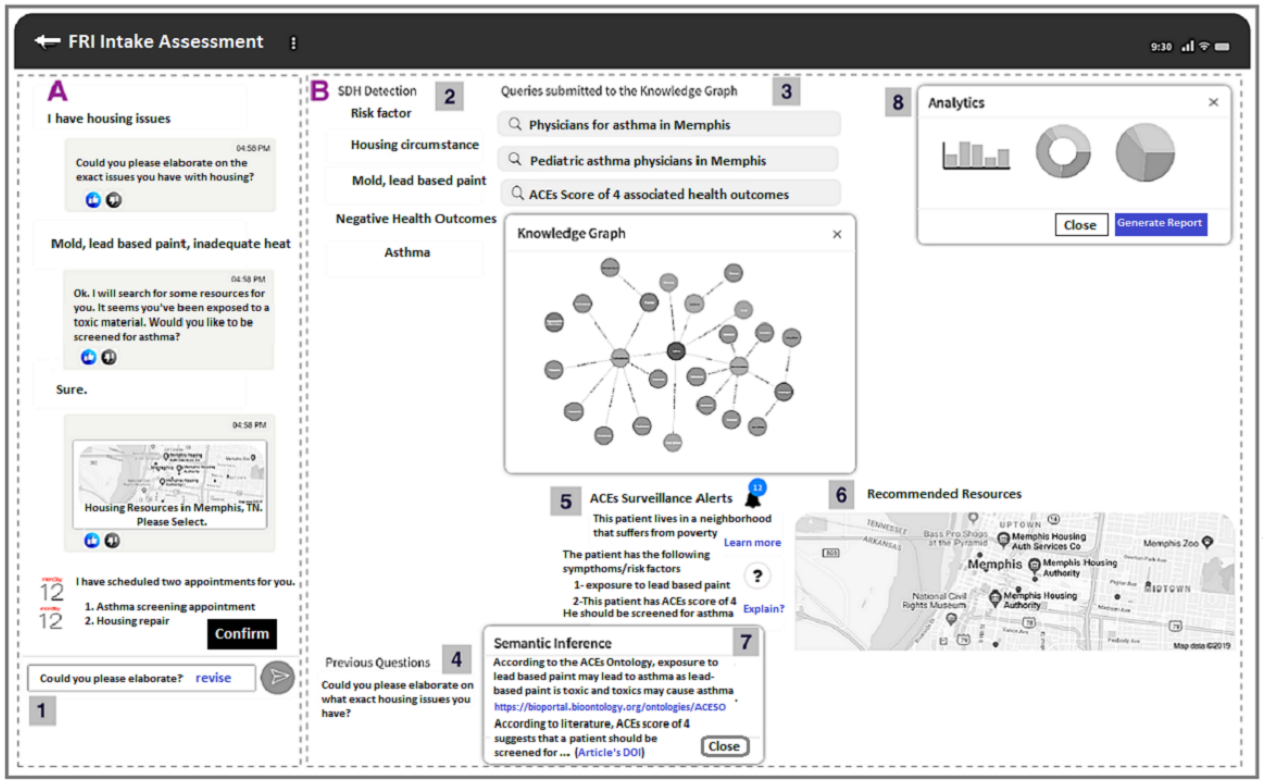
The aggregated results from all user conversations provide policymakers with insights about population health. Analyzing such data can help in designing interventions and preventive measures based on the most prevalent risk factors in certain regions. It can also assist the framework users by providing recommendations on how to direct future conversations. For example, it can look into smaller communities within geographic areas or perform collaborative filtering based on similarities in user behavior, or detect similar communities on the conversation graph. For example, if a user that belongs to a certain age or ethnicity group shows a certain pattern/route during the case assessment conversation procedure, then the QA agent may suggest the same route for the next user with similar criteria.

Results

We describe the main features provided by SPACES through a proof-of-concept prototype that will render the information collected by the QA agent and the recommendation service on a user-friendly interface. The prototype is intended for several types of users, including caregivers (eg, child-parents) or health care professionals (eg, nurses, physicians, social workers). The main features of the view that would be available to a health care professional are illustrated in Figure 7, including recommendations for digital assistance (A), studying the association between ACEs and SDoH (B.2 and B.5), knowledge graph querying (B.3), geocoded resource recommendation (B.6), and explainability by displaying inference sources (B.5 and B.7). The QA agent view that would be available to a caregiver is illustrated in Figure 5 and in panel A of Figure 7.

We present multiple use-case scenarios in Textbox 3. For simplicity, we use Scenario 4 to demonstrate how the QA agent detects context and lays it over to the recommendation service, and how the recommendation service uses contextual parameters to generate a personalized knowledge graph.

Figure 7. Prototype of the recommendation system. (A) Question-answering (QA) view (1) send/revise suggested questions. (B) Health care practitioner assessment panel, including (2) social determinants of health (SDoH) detection, (3) knowledge graph and queries, (4) pool of previously asked questions,(5) alerts for detected adverse childhood experiences (ACEs) symptoms, (6) geocoded resource allocation, (7) explain recommendations, (8) and visual analytics.



Textbox 3. Four use case scenarios.

Scenario 1: Assessing needs relating to SDoH. SDoH: social determinants of health; N/A: not applicable; ACEs: adverse childhood experiences; UPHO: Urban Population Health Observatory.

"I am currently residing in a safe place, but I'm concerned about my household income as I am currently unemployed due to legal issues. I have some college and I am interested in furthering my education, but would prefer a job first."

Symptoms: N/A

Risk factors: ACEs detected: Unemployed due to legal issues (Legal, Textbox 1)

I'm interested in furthering my education (Education, Textbox 1)

Outcomes: N/A

Intervention: Information on how to get into the school

Follow up on legal issues.

Scenario 2: Assessing issues relating to ACEs

"My husband is an alcoholic and he had served time in jail and right now it is hard to soothe my 4-year-old baby boy or calm him down. He also bounces back quickly when things do not go his way. This just puts a lot of pressure on me"

Symptoms: Hard to soothe or calm down (behavioral)

Bounces back quickly (behavioral)

Risk factors: ACEs detected: living in a household with substance abuse

ACEs detected: living with a household member who was in jail

Outcomes: provided (stress)

inferred (asthma)

Intervention: Therapeutic (schedule psychologist appointment)

Scenario 3: Mixed model of ACEs and SDoH

"I have a couple of issues. My 7-year-old son is developmentally delayed, and we have food insecurities that we hope we could resolve before the holidays. But I am mostly concerned about food."

Symptoms: Child developmental delay (behavioral)

Risk factors: food insecurities

Outcomes: N/A

Intervention: Provide information about food pantries

Schedule psychologist appointment

Scenario 4: Detecting risk factors for potential negative health outcomes and providing early diagnosis

"I am a Hispanic 21-year-old female living in Memphis. My 6-year-old child experiences night terror. I have recently separated from my husband."

Symptoms: Night terror (emotional neglect causes night terror)

Risk factors: ACEs detected:

provided (living in a household of divorce)

inferred (ontology, emotional neglect)

(UPHO, a neighborhood with blight)

Outcomes: inferred (asthma)

Intervention: Therapeutic (schedule a medical appointment)

Based on Scenario 4, Alice is a Hispanic female located in a city with zip code 38103 and has a 6-year-old child who suffers from night terrors. Alice can either provide vague (Figure 5a) or detailed (Figure 5b) text. Either way, the agent collects as many contextual parameters as possible and then passes them over to the recommendation service.

The service starts building the personalized graph for Alice by adding a node labeled *zip code* and *has_value* 38103. It then links the node to all concepts related to a zip code, including neighborhood and physical characteristics. The physical characteristics are then linked to all related concepts based on the ACESO, including blight, walkability, and crime, and it populates these nodes with values from the UPHO. If blight has the maximum value, it links the blight concept with related concepts based on the axiom (blight causes exposure to toxicants and exposure to toxicants leads to asthma). Thus, it links it to the toxicant and asthma nodes. Additionally, it adds the age

node and populates it with a contextual parameter value of 6, and links the age concept to the school-age concept. It then adds the *symptom* node and links it to *night terror* based on the concept subtype relation in the ontology. It then links *night terror* to *emotional neglect* based on an axiom. The inference path in Alice's scenario is shown on the graph in red in Figure 6 to illustrate how Alice's case leads to a recommendation for pediatric asthma physicians. The metapath, derived from that concrete path, can be used in future recommendations if an individual is encountered with similar contextual parameters.

The resulting graph is a property graph in that both nodes and relations have properties. For instance, the inferred relations (eg, *shouldBeScreenedFor* asthma, has an ACEs score of 4, and *livesIn* a neighborhood with 80% blight prevalence) can be labeled with the source of inference, confidence probability, severity, and type of intervention. Sources of knowledge inference appear in yellow boxes at the bottom of Figure 6, which include UPHO, research papers, the ACESO, and others. We can enrich the graph further with a clinic locator graph of physicians who can provide prescriptions for pediatric asthma, and then filter the resource nodes further based on provided contextual parameters. For instance, we can keep only clinics that fall within a small radius from Alice's residential zip code and that provide Spanish language–speaking services based on her ethnicity.

The health care practitioner can observe the inferred knowledge through several features rendered on the prototype interface as follows: (1) ACEs alerts (eg, Alice's case indicates an ACEs score of 4, night terror as a symptom, and suggests an early diagnosis of asthma as an outcome) (Figure 7, B.2); (2) SDoH alerts (eg, the risk factors associated with Alice's case) (Figure 7, B.5); (3) visual aids for explainability by clicking on the (?) icons to display links to inference sources (eg, the ontology link on BioPortal and the identifier for the CDC-Kaiser paper) (Figure 7, B.7); (4) geocoded view of the suggested resources (eg, clinics in Memphis) (Figure 7, B.6); and (5) which questions to ask next (Figure 7, B.4).

Discussion

Principal Findings

The significance of the proposed approach lies in its ability to provide recommendations to the QA agent with the least effort from both the user and the health care practitioner. It aims to maximize knowledge about the patient without having to delve into all of the questions that are often asked in ACEs and SDoH intake assessments. It also provides the ability to explain why a certain question or resource was suggested.

Preliminary rapid prototyping of the recommender system allowed for early verification of the functionalities through the multiple case scenarios described in this paper. We anticipate rapid feedback from end users on various features during an iterative development process, and finally establish a comprehensive usability and user experience test. We have evaluated the SPACES semantic framework and its underlying ontology automatically using description logic reasoners such as Fact++ [20] to ensure the consistency and satisfiability of the ontology and semantic model. We also used inputs from our collaborating domain experts to assess the soundness and completeness of the ontology by examining how well it has been aligned with the required criteria defined in our domain and scope. Furthermore, we evaluated the usability of ACESO based on its functionality and capability to respond and answer the target queries. Finally, further work is underway to develop and conduct a series of formal evaluation practices for a comprehensive assessment of the accuracy, usability, coverage, confidence, trustability, as well as adaptivity and scalability of the recommender system [21]. Moreover, we will assess the utility and impact of the system on the surveillance of ACEs to generate a timely response and intervention, ultimately informing public health planning and policymaking.

The proposed approach might face some limitations. One limitation is in providing an overall guiding architecture to support transfer ability between health domains. Several of the QA platforms (eg, Google DialogFlow and IBM Watson) read rules on how to answer questions from backend sources (eg, HTML FAQ files, Plaintext files). These sources can help load the questions into the QA agent. They also require training the agent with concepts that may appear in the QA text. The ontology in our case helps load the concepts automatically, which is separate from the QA platform implementation itself. Each domain will have its own concepts that can be encoded in a separate ontology, and we can either develop new ontologies or reuse existing ones.

Another limitation is that the recommendation system has access to only population-wide data, where the population's characteristics might be different than the characteristics of the individuals living in that population or neighborhood. However, for specific users and specific use cases, it needs access to individual data. For instance, a pediatrician trying to decide whether a child is suffering from ACEs will need to have access to the child's health history and other relevant information but should not be allowed to access information about the parent's finances or criminal history (outside of what is publicly available). Moreover, a local judge will want to have access to the criminal history of the family members if they want to decide whether the child should be removed from their parents to ensure their safety, but they should not be able to access any of their medical records. Thinking about how to control mediated access to sensitive information will be a key part of the development of the recommendation system. We are currently working on integrating the recommendation system into a personalized health library [22].

The adoption of recommendation systems may be hindered by a poor user-interface design or poor integration into clinical workflows. Human factors engineering can improve efficiency, reduce errors, increase technology adoption, and reduce the early abandonment of systems. This paper lacks an objective evaluation of *how* health care practitioners will benefit from the explanations or how the quality of those explanations would be assessed. We plan to use an iterative user-centered design and formative evaluation by conducting predevelopment focus groups, which might reveal issues related to manual data entry from target users and the time spent reviewing generated knowledge from providers. Ongoing research on this project

will also involve implementing an optimization algorithm of the recommendations. We will also consider technically evaluating the system for precision and performance.

All updates in the underlying ontologies and semantic structure will be managed through our previously implemented framework [23, 24]. As for the target audience, the system is intended for a variety of users depending on the domain, including social workers, health care providers (eg, nurses, physicians), and caregivers (child-parent). The digital assistant or QA agent is the part intended and available for end users (eg, caregivers). It is a simulation of the conversations that occur between health care providers or social workers and caregivers as they discuss their case. The text exchanged through the chat is used in the backend for both enhancing the QA agent and refining the personal knowledge graph of the current user. The visual graphs and underlying reasoning are intended for AI explainability, which is critical for health care providers to make decisions. In particular, the analytics part is intended for monitoring and research purposes, and therefore is more appropriate for health professionals and policymakers.

Finally, we discuss the implications for policy and practice. We believe that early intervention is the best way to prevent the progression of negative health outcomes to their end stage, and that well-designed early detection systems can aid clinicians by generating knowledge that can be aligned with clinical workflows. Thus, systems tailored toward such interventions by utilizing knowledge about self-reported or detected SDoH and ACEs would be most useful. Through the framework, decision makers can (1) identify risk factors, (2) integrate and validate ACEs and SDoH exposure at individual and population levels, and (3) detect high-risk groups. The analytics component could be most useful for policymakers for this purpose and is intended for monitoring and research purposes.

Conclusions

In this study, we leveraged explainable AI to present a proof-of-concept prototype for a knowledge-driven evidence-based recommendation system to improve the surveillance of ACEs. The proposed approach will enhance the health care practitioner's ability to provide explanations for the decisions that they make. Further development and official evaluation are underway to include a privacy layer through a personal health library and to conduct a clinical trial for formally assessing both the usability and usefulness of the implementation.

Acknowledgments

We would like to thank Dr Robert L Davis, Dr Jonathan A McCullers, Dr Jason Yaun, Dr Sandra R Arnold, and the entire team at the Family Resilience Initiative at Le Bonheur Children's Hospital, Memphis, Tennessee, for their support and insights. This research was partially supported by the Memphis Research Consortium.

Conflicts of Interest

None declared.

References

1. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The Adverse Childhood Experiences (ACE) Study. Am J Prev Med 1998 May;14(4):245-258. [doi: 10.1016/s0749-3797(98)00017-8] [Medline: 9635069]
2. Social Determinants of Health. World Health Organization. 2019. URL: https://www.who.int/social_determinants/sdh_definition/en/ [accessed 2020-09-17]
3. Shin EK, Shaban-Nejad A. Urban Decay and Pediatric Asthma Prevalence in Memphis, Tennessee: Urban Data Integration for Efficient Population Health Surveillance. IEEE Access 2018;6:46281-46289. [doi: 10.1109/access.2018.2866069]
4. Chung EK, Siegel BS, Garg A, Conroy K, Gross RS, Long DA, et al. Screening for Social Determinants of Health Among Children and Families Living in Poverty: A Guide for Clinicians. Curr Probl Pediatr Adolesc Health Care 2016 May;46(5):135-153 [FREE Full text] [doi: 10.1016/j.cppeds.2016.02.004] [Medline: 27101890]
5. Shorey S, Ang E, Yap J, Ng ED, Lau ST, Chui CK. A Virtual Counseling Application Using Artificial Intelligence for Communication Skills Training in Nursing Education: Development Study. J Med Internet Res 2019 Oct 29;21(10):e14658 [FREE Full text] [doi: 10.2196/14658] [Medline: 31663857]
6. Battaglia P, Hamrick J, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint 2018 Oct 17 [FREE Full text]
7. Ehrlinger L, Wöß W. Towards a Definition of Knowledge Graphs. In: SEMANTiCS (Posters, Demos, SuCCESS). 2016 Presented at: 12th International Conference on Semantic Systems; September 12-15, 2016; Leipzig, Germany.
8. Ruan T, Huang Y, Liu X, Xia Y, Gao J. QAnalysis: a question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research. BMC Med Inform Decis Mak 2019 Apr 01;19(1):82 [FREE Full text] [doi: 10.1186/s12911-019-0798-8] [Medline: 30935389]
9. Goodwin TR, Harabagiu SM. Medical Question Answering for Clinical Decision Support. Proc ACM Int Conf Inf Knowl Manag 2016 Oct;2016:297-306 [FREE Full text] [doi: 10.1145/2983323.2983819] [Medline: 28758046]
10. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a Health Knowledge Graph from Electronic Medical Records. Sci Rep 2017 Jul 20;7(1):5994. [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]

11. Nelson CA, Butte AJ, Baranzini SE. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. Nat Commun 2019 Jul 10;10(1):3045. [doi: 10.1038/s41467-019-11069-0] [Medline: 31292438]
12. Brenas JH, Shin EK, Shaban-Nejad A. Adverse Childhood Experiences Ontology for Mental Health Surveillance, Research, and Evaluation: Advanced Knowledge Representation and Semantic Web Techniques. JMIR Ment Health 2019 May 21;6(5):e13498 [FREE Full text] [doi: 10.2196/13498] [Medline: 31115344]
13. Brenas JH, Shin EK, Shaban-Nejad A. An Ontological Framework to Improve Surveillance of Adverse Childhood Experiences (ACEs). Stud Health Technol Inform 2019;258:31-35. [Medline: 30942708]
14. Brenas JH, Shin EK, Shaban-Nejad A. A Hybrid Recommender System to Guide Assessment and Surveillance of Adverse Childhood Experiences. Stud Health Technol Inform 2019 Jul 04;262:332-335. [doi: 10.3233/SHTI190086] [Medline: 31349335]
15. Adverse Childhood Experiences Ontology. BioPortal. 2019 Feb 25. URL: https://bioportal.bioontology.org/ontologies/ACESO [accessed 2020-09-17]
16. Shin EK, Kwon Y, Shaban-Nejad A. Geo-clustered chronic affinity: pathways from socio-economic disadvantages to health disparities. JAMIA Open 2019 Oct;2(3):317-322 [FREE Full text] [doi: 10.1093/jamiaopen/ooz029] [Medline: 31984364]
17. Dialogflow. Google Cloud. 2010. URL: https://dialogflow.com/ [accessed 2022-10-20]
18. Chang X, Jiang X, Mkandarwire T, Shen M. Associations between adverse childhood experiences and health outcomes in adults aged 18-59 years. PLoS One 2019;14(2):e0211850 [FREE Full text] [doi: 10.1371/journal.pone.0211850] [Medline: 30730980]
19. Neo4j Platform. URL: https://neo4j.com/ [accessed 2020-09-17]
20. Tsarkov D, Horrocks I. FaCT++ Description Logic Reasoner: System Description. In: Furbach U, Shankar N, editors. Automated Reasoning. IJCAR 2006. Lecture Notes in Computer Science, vol 4130. 2006 Presented at: International Joint Conference on Automated Reasoning; August 17-20, 2006; Seattle, WA URL: https://link.springer.com/chapter/10.1007%2F11814771_26 [doi: 10.1007/11814771_26]
21. Shani G, Gunawardana A. Evaluating Recommendation Systems. In: Ricci F, Rokach L, Shapira B, Kantor P, editors. Recommender Systems Handbook. Boston: Springer; 2011:257-297.
22. Ammar N, Bailey JE, Davis RL, Shaban-Nejad A. The Personal Health Library: A Single Point of Secure Access to Patient Digital Health Information. Stud Health Technol Inform 2020 Jun 16;270:448-452. [doi: 10.3233/SHTI200200] [Medline: 32570424]
23. Shaban-Nejad A, Haarslev V. Managing changes in distributed biomedical ontologies using hierarchical distributed graph transformation. Int J Data Min Bioinform 2015;11(1):53-83. [doi: 10.1504/ijdmb.2015.066334] [Medline: 26255376]
24. Shaban-Nejad A, Ormandjieva O, Kassab M, Haarslev V. Managing Requirement Volatility in an Ontology-Driven Clinical LIMS Using Category Theory. Int J Telemed Appl 2009;2009:917826. [doi: 10.1155/2009/917826] [Medline: 19343191]

Abbreviations

ACEs: adverse childhood experiences
ACESO: Adverse Childhood Experiences Ontology
AI: artificial intelligence
API: application programming interface
FRI: Family Resilience Initiative
ML: machine learning
QA: question-answering
RDF: Resource Description Framework
SDoH: social determinants of health
SPACES: Semantic Platform for Adverse Childhood Experiences Surveillance
UPHO: Urban Population Health Observatory

Edited by J Bian; submitted 16.03.20; peer-reviewed by SA Chun, M Mulvenna; comments to author 22.06.20; revised version received 25.08.20; accepted 08.10.20; published 04.11.20.

Please cite as:
Ammar N, Shaban-Nejad A
Explainable Artificial Intelligence Recommendation System by Leveraging the Semantics of Adverse Childhood Experiences: Proof-of-Concept Prototype Development
JMIR Med Inform 2020;8(11):e18752
URL: https://medinform.jmir.org/2020/11/e18752/
doi:10.2196/18752
PMID:33146623

Original Paper

Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models

Xi Yang[1], PhD; Xing He[1], MSc; Hansi Zhang[1], MSc; Yinghan Ma[1], BSc; Jiang Bian[1], PhD; Yonghui Wu[1], PhD

Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States

Corresponding Author:
Yonghui Wu, PhD
Department of Health Outcomes and Biomedical Informatics
University of Florida
2004 Mowry Road
Gainesville, FL, 32610
United States
Phone: 1 352 294 8436
Email: yonghui.wu@ufl.edu

Abstract

Background: Semantic textual similarity (STS) is one of the fundamental tasks in natural language processing (NLP). Many shared tasks and corpora for STS have been organized and curated in the general English domain; however, such resources are limited in the biomedical domain. In 2019, the National NLP Clinical Challenges (n2c2) challenge developed a comprehensive clinical STS dataset and organized a community effort to solicit state-of-the-art solutions for clinical STS.

Objective: This study presents our transformer-based clinical STS models developed during this challenge as well as new models we explored after the challenge. This project is part of the 2019 n2c2/Open Health NLP shared task on clinical STS.

Methods: In this study, we explored 3 transformer-based models for clinical STS: Bidirectional Encoder Representations from Transformers (BERT), XLNet, and Robustly optimized BERT approach (RoBERTa). We examined transformer models pretrained using both general English text and clinical text. We also explored using a general English STS dataset as a supplementary corpus in addition to the clinical training set developed in this challenge. Furthermore, we investigated various ensemble methods to combine different transformer models.

Results: Our best submission based on the XLNet model achieved the third-best performance (Pearson correlation of 0.8864) in this challenge. After the challenge, we further explored other transformer models and improved the performance to 0.9065 using a RoBERTa model, which outperformed the best-performing system developed in this challenge (Pearson correlation of 0.9010).

Conclusions: This study demonstrated the efficiency of utilizing transformer-based models to measure semantic similarity for clinical text. Our models can be applied to clinical applications such as clinical text deduplication and summarization.

(JMIR Med Inform 2020;8(11):e19735) doi:10.2196/19735

KEYWORDS

clinical semantic textual similarity; deep learning; natural language processing; transformers

Introduction

Semantic textual similarity (STS) is a natural language processing (NLP) task to quantitatively assess the semantic similarity between two text snippets. STS is usually approached as a regression task where a real-value score is used to quantify the similarity between two text snippets. STS is a fundamental NLP task for many text-related applications, including text deduplication, paraphrasing detection, semantic searching, and question answering. In the general English domain, semantic evaluation (SemEval) STS shared tasks have been organized annually from 2012 to 2017 [1-6], and STS benchmark datasets were developed for evaluation [6]. Previous work on STS often used machine learning models [7-9] such as support vector machine [10], random forest [11], convolutional neural networks [12], and recurrent neural networks [13] and topic modeling techniques [8] such as latent semantic analysis [14] and latent Dirichlet allocation [15]. Recently, deep learning models based on transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT) [16], XLNet [17], and Robustly optimized BERT approach (RoBERTa) [18] have

demonstrated state-of-the-art performances on the STS benchmark dataset [19] and remarkably outperformed the previous models. More recently, the Text-to-Text Transfer Transformer model [20] and the StructBERT model [21] have further improved the performance on the STS benchmark. These studies demonstrated the efficiency of transformer-based models for STS tasks.

Rapid adoption of electronic health record (EHR) systems has made longitudinal health information of patients available electronically [22,23]. EHRs consist of structured, coded data and clinical narratives. The structured EHR data are typically stored as predefined medical codes (eg, International Classification of Diseases, 9th/10th Revision, codes for diagnoses) in relational databases. Various common data models were used to standardize EHR data to facilitate downstream research and clinical studies [24]. However, clinical narratives are often documented in a free-text format, which contains many types of detailed patient information, such as family history, adverse drug events, and medical imaging result interpretations, that are not well captured in the structured medical codes [25]. As free text, the clinical notes may contain a considerable amount of duplication, error, and incompleteness for various reasons (eg, copy-and-paste or using templates and inconsistent modifications) [26,27]. STS can be applied to assess the quality of the clinical notes and reduce redundancy to support downstream NLP tasks [28]. However, up until now, only a few studies [29-31] have explored STS in the clinical domain due to the limited data resources for developing and benchmarking clinical STS tasks. Recently, a team at the Mayo Clinic developed a clinical STS dataset, MedSTS [32], which consists of more than 1000 annotated sentence pairs extracted from clinical notes. Based on the MedSTS dataset, the 2018 BioCreative/Open Health NLP (OHNLP) challenge [33] was organized as the first shared task examining advanced NLP methods for STS in the clinical domain. In this challenge, two different teams explored various machine learning approaches, including several deep learning models [30,31]. Later, more teams competed in the 2019 National NLP Clinical Challenges (n2c2)/OHNLP STS challenge with a larger clinical STS dataset [34]. During this challenge, many new emerging NLP techniques, such as transformer-based models, were explored.

This study presents our machine learning models developed for the 2019 n2c2/OHNLP STS challenge. We explored state-of-the-art transformer-based models (BERT, XLNet, and RoBERTa) for clinical STS. We systematically examined transformer models pretrained using general English corpora and compared them with clinical transformer models pretrained using clinical corpora. We also proposed a representation fusion method to ensemble the transformer-based models. In this challenge, our clinical STS system based on the XLNet model achieved a Pearson correlation score of 0.8864, ranked as the third-best performance among all participants. After the challenge, we further explored a new transformer-based model, RoBERTa, which improved the performance to 0.9065 and outperformed the best performance (0.9010) reported in this challenge. This study demonstrated the efficiency of transformer-based models for STS in the clinical domain.

Methods

Dataset

The 2019 n2c2 organizers developed a corpus of 2054 sentence pairs derived from over 300 million deidentified clinical notes from the Mayo Clinic's EHR data warehouse. The sentence pairs were divided into a training set of 1642 sentence pairs for model development and a test set of 412 sentence pairs for evaluation. Similar to the annotation scheme in the general English domain, the challenge corpus was annotated by assigning a similarity score for each sentence pair as a number on a scale from 0.0 to 5.0, where 0.0 indicates that the semantics of the two sentences are entirely independent (ie, no overlap in their meanings), and 5.0 signifies that two sentences are semantically equivalent. Annotators used arbitrary similarity scores between 0.0 and 5.0, such as 2.5 or 3.5, to reflect different levels of equality. Table 1 presents the descriptive statistics of the datasets. The distribution of similarity scores is quite different between the training and test datasets. In the training set, the range with the most cases (509/1642, 31.0%) was (3.0, 4.0], whereas in the test set, most scores (238/412, 57.8%) were distributed in the range (0.0, 1.0]. In this study, we denoted this challenge dataset as STS-Clinic. In addition to the STS-Clinic, we also used a general English domain STS benchmark dataset from the SemEval 2017 [6] as an external source. We merged the original training and development datasets to create a unique dataset of 7249 annotated sentence pairs. We denoted this combined general English domain dataset as STS-General and used it as a complementary training set for model development in this study. Compared to the STS-Clinic, the similarity scores in STS-General were more evenly distributed in different ranges (Table 1).

Table 1. Descriptive statistics of the datasets.

Dataset	Sentence pairs, n	Annotation distribution, n (%)				
		[0.0, 1.0]	(1.0, 2.0]	(2.0, 3.0]	(3.0, 4.0]	(4.0, 5.0]
STS-Clinic[a] Training	1642	312 (19.0)	154 (9.4)	394 (24.0)	509 (31.0)	273 (16.6)
STS-Clinic Test	412	238 (57.8)	46 (11.2)	32 (7.8)	62 (15.0)	34 (8.3)
STS-General Training	7249	1492 (20.6)	1122 (15.5)	1413 (19.5)	1260 (17.4)	1962 (27.1)

[a]STS: semantic textual similarity.

Preprocessing of Sentence Pairs

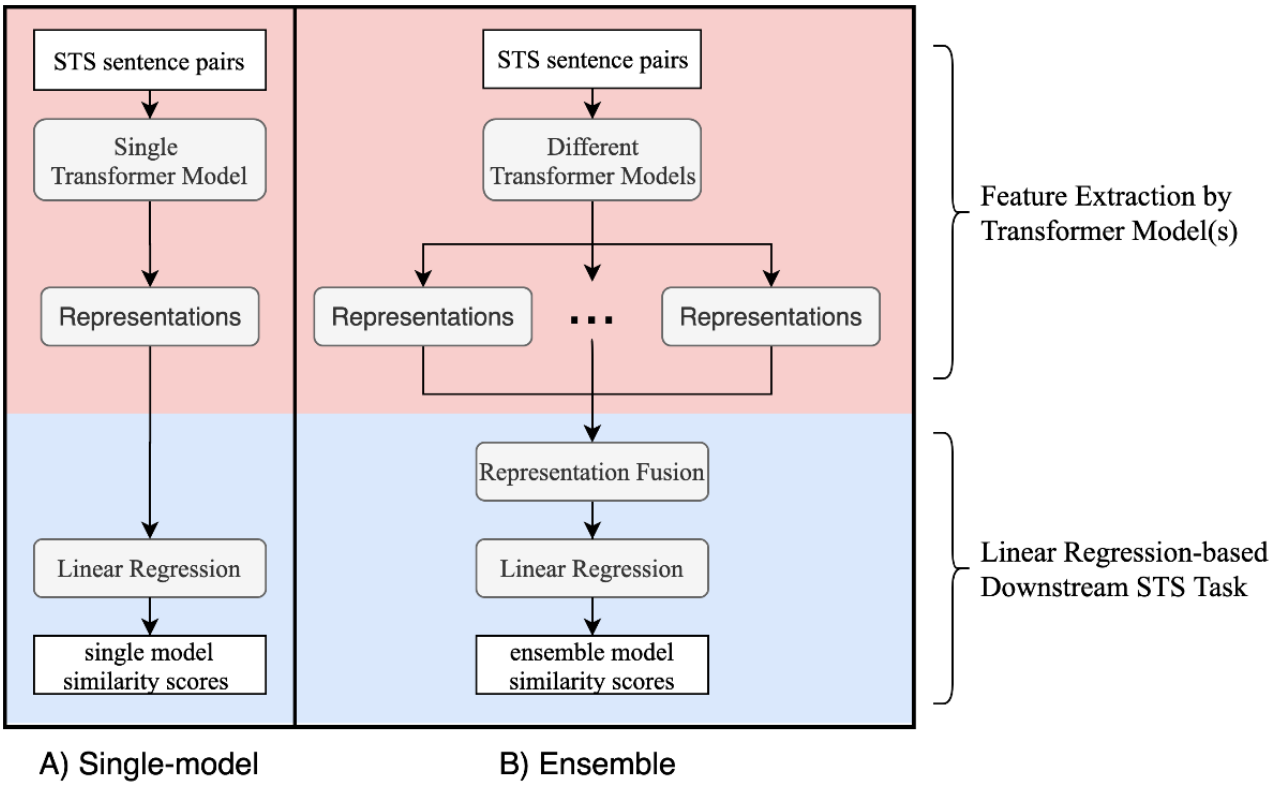
We developed a preprocessing pipeline to normalize each sentence pair, including (1) converting all words to lower case; (2) inserting white spaces to separate words from punctuation (eg, "[ab/cd]" → "[ab / cd]"; "abc,def" → "abc , def"); and (3) replacing two or more spaces or tabs ("\t") with a single space. We did not remove any stop-words from the sentences and kept the original formats of the numbers without any conversion. Since different transformer models adopted different tokenization strategies (eg, WordPiece for BERT, byte pair encoding for RoBERTa, and SentencePiece for XLNet), our preprocessing automatically picked the appropriate tokenizer according to the transformer model in use.

Transformer Model-Based STS System

In this study, we investigated three transformer models (BERT, XLNet, and RoBERTa) for clinical STS. BERT is a bidirectional transformer-based encoder model pretrained with a combination of masked language modeling (MLM) and next sentence prediction. RoBERTa has the same architecture as BERT but pretrained with a robust optimizing strategy. The RoBERTa pretraining procedure used dynamic MLM but removed the next sentence prediction task. XLNet is a transformer-based model pretrained with the bidirectional autoregressive language modeling method. Unlike the MLM used by BERT and RoBERTa, the autoregressive language model uses data permutation instead of data corruption and reconstruction. All three transformer models provided two different settings: a "BASE" setting and a "LARGE" setting. The main difference between the BASE model and the LARGE model is the number of layers. For example, the BERT-base model features 12 layers of transformer encoder layers, 768 hidden units in each layer, and 12 attention heads, while the BERT-large consists of 24 transformer blocks with a hidden size of 1024 and 16 attention heads. The total number of parameters for the BERT-large model is approximately 340 million, which is about 3 times more than the BERT-base model. In this study, we explored general transformers (pretrained using general English corpora) using both the BASE model and the LARGE model. We also examined clinical transformers pretrained using clinical notes from the MIMIC-III database. For clinical transformers, we adopted the BASE settings as we did not observe additional benefits from using the LARGE setting.

As shown in Figure 1, our STS system has two modules: (1) a transformer model–based feature learning module and (2) a regression-based similarity score learning module. In the feature learning module, transformer-based models were applied to learn distributed sentence-level representations from sentence pairs. In the similarity score learning module, we adopted a linear regression layer to calculate a similarity score between 0.0 and 5.0 according to the distributed representations derived from the transformers. We explored both single-model and ensemble solutions. Figure 1A shows the single-model solution where only one transformer-based model was used for feature representation learning. Figure 1B shows the ensemble solution where different transformer models were integrated. Ensemble learning is an efficient approach to aggregate different machine learning models to achieve better performance [35]. In this work, we tried different strategies to combine the distributed representations from two or three transformers as a new input layer for the similarity score learning module. We explored several methods to combine the distributed representations from different transformers, including (1) simple head-to-tail concatenation, (2) pooling, and (3) convolution.

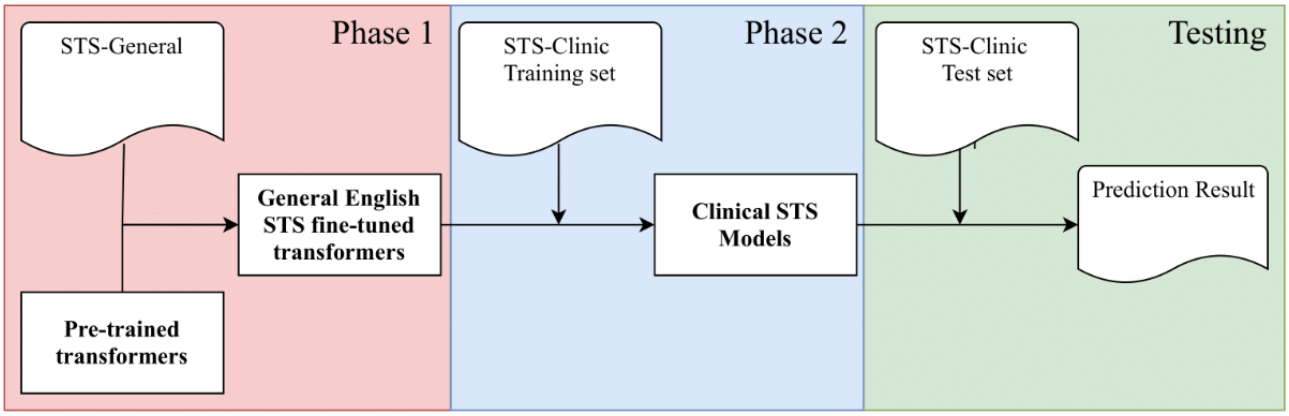
Figure 1. An overview of our single-model and ensemble solutions for clinical STS. STS: semantic textual similarity.



Training Strategy

As shown in Figure 2, we adopted a two-phase procedure to train our clinical STS models. In the first phase, an intermediate STS model was fine-tuned using the STS-General corpus. Subsequently, the intermediate model was further fine-tuned using the STS-Clinic corpus in phase 2. The fine-tuned model from the second phase was used for final testing. We used 5-fold cross-validation for hyperparameter optimization in both phase 1 and phase 2 training. We optimized the epoch number, batch size, and learning rate according to the cross-validation results.

Figure 2. The two-stage procedure for clinical STS model development. STS: semantic textual similarity.



Experiments and Evaluations

In this study, we implemented our STS system using the Transformers library developed by the HuggingFace team [36]. We also used the PyTorch-based general transformer models trained using general English corpora maintained by the HuggingFace team. The clinical transformer models were derived by further pretraining these general transformer models with clinical notes from the MIMIC-III database [37]. Table 2 shows the hyperparameters used for each transformer model. For evaluation, the results were calculated as the Pearson correlation scores using the official evaluation script provided by the 2019 n2c2/OHNLP challenge organizers. To report the *P* value for each Pearson correlation score, we adopted the SciPy package [38].

Table 2. Hyperparameters for transformer models.

Model	Number of epochs	Batch size	Learning rate[a]
BERT-base[b]	4	8	1.00E-05
BERT-mimic	3	8	1.00E-05
BERT-large	3	8	1.00E-05
XLNet-base	3	4	1.00E-05
XLNet-mimic	3	4	1.00E-05
XLNet-large	4	4	1.00E-05
RoBERTa-base[c]	3	4	1.00E-05
RoBERTa-mimic	3	4	1.00E-05
RoBERTa-large	3	4	1.00E-05
BERT-large + XLNet-large	4	8	1.00E-05
BERT-large + RoBERTa-large	3	4	1.00E-05
RoBERTa-large + XLNet-large	4	4	1.00E-05
BERT-large + XLNet-large + RoBERTa-large	3	2	1.00E-05

[a]The learning rate is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function [39].

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]RoBERTa: Robustly optimized BERT approach.

Results

Table 3 compares the performance of the different transformer models on the test dataset. The RoBERTa-large model achieved the best Pearson correlation of 0.9065 among all models, which outperformed the two models we developed and submitted during the challenge, including the XLNet-large (a Pearson correlation score of 0.8864) and the BERT-large models (a

Pearson correlation score of 0.8549). For RoBERTa and XLNet, the models developed using the LARGE setting pretrained using general English corpora achieved better performances than their BASE settings (0.9065 vs 0.8778 for RoBERTa; 0.8864 vs 0.8470 for XLNet, respectively), whereas the BERT-base achieved a Pearson correlation score of 0.8615 that outperformed the BERT-large model's score of 0.8549. For all transformers, the models pretrained using general English corpora (in both LARGE settings and BASE settings) outperformed their corresponding clinical models pretrained using clinical notes from the MIMIC-III database. Among the ensemble models, the BERT-large + RoBERTa-large model achieved the best Pearson correlation score of 0.8914, which is remarkably lower than the best model, RoBERTa-large. We also observed that the performances of ensemble models were often in between the two individual models (eg, BERT-large + RoBERTa-large achieved 0.8914, which is between the BERT-large score of 0.8549 and RoBERTa-large score of 0.9065). The ensemble model of all three transformers achieved a Pearson correlation of 0.8452, which was even worse.

Table 3. Performances of the Pearson correlation on the test set.

Model	Pearson correlation on test set	*P* value
BERT-base[a]	0.8615	<.001
BERT-mimic	0.8521	<.001
BERT-large[b]	0.8549	<.001
XLNet-base	0.8470	<.001
XLNet-mimic	0.8286	<.001
XLNet-large[b,c]	0.8864	<.001
RoBERTa-base[d]	0.8778	<.001
RoBERTa-mimic	0.8705	<.001
RoBERTa-large	*0.9065*	<.001
BERT-large + XLNet-large[b]	0.8764	<.001
BERT-large + RoBERTa-large	0.8914	<.001
RoBERTa-large + XLNet-large	0.8854	<.001
BERT-large + XLNet-large + RoBERTa-large	0.8452	<.001

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]The challenge submissions.

[c]The best challenge submission (ranked 3rd).

[d]RoBERTa: Robustly optimized BERT approach.

Discussion

Principal Results

Clinical STS is a fundamental task in biomedical NLP. The 2019 n2c2/OHNLP shared task was organized to solicit state-of-the-art STS algorithms in the clinical domain. We participated in this challenge and developed a deep learning–based system using transformer-based models. Our best submission (XLNet-large) achieved the third-best performance (a Pearson correlation score of 0.8864) among the 33 teams. Based on our participation, we further explored RoBERTa models and improved the performance to 0.9065 (RoBERTa-large), demonstrating the efficiency of transformer models for clinical STS. We also further explored three different ensemble strategies to develop ensembled models using transformers. Our experimental results show that the ensemble methods did not outperform the unified individual models. Another interesting finding is that the transformers pretrained using the clinical notes from the MIMIC-III database did not outperform the general transformers pretrained using general English corpora on clinical STS. One possible reason might be that the clinical corpora we used for training are relatively small compared with the general English corpus. Further investigation examining these findings is warranted.

Experiment Findings

Although previous studies [40-44] have shown that pretraining transformer models with domain-specific corpora could enhance their performances in domain-related downstream tasks (such as clinical concept extraction), our results in this study indicated that this strategy might not be helpful for clinical STS. For all three types of transformers explored in this study, the models pretrained using general English text consistently obtained higher scores than the corresponding models pretrained using clinical text. For example, the Pearson correlation score achieved by the RoBERTa-mimic was 0.8705; however, the RoBERTa-base yielded a higher performance of 0.8778. Tawfik et al [45] have similarly observed that the PubMed pretrained BioBERT did not outperform the corresponding general BERT model pretrained using English text on clinical STS.

In the clinical STS task, using STS-General (an STS corpus annotated in the general English domain) as an extra training

set in addition to STS-Clinic could efficiently improve performances for transformer-based models. Taking the RoBERTa model as an example, the RoBERTa-large fine-tuned using only the clinical text (ie, STS-Clinic) achieved a Pearson correlation score of 0.8720; however, the same model fine-tuned with both the general English text (ie, STS-General) and clinical text (ie, STS-Clinic) achieved a score of 0.9065 (approximately 0.035 higher). We observed similar results for BERT and XLNet. Without Phase 1 (Figure 2), the BERT-large and XLNet-large models achieved Pearson correlation scores of 0.8413 and 0.8626, respectively, which are lower than the results we submitted (0.8549 and 0.8864) using two-phase training. We looked into the training datasets for possible reasons. Although the STS-General and STS-Clinic were extracted from different domains, there are common contents shared between them. First, the annotation guidelines between the two datasets were highly aligned. For both datasets, the annotation scale is from 0.0 to 5.0, and each score reflects the same similarity level. Since the two STS datasets were annotated by different annotators, subjective annotation bias might be introduced (eg, the judgement and agreement of semantic similarity among annotators might be different in the two datasets). However, our experiment results showed that training with both datasets improved the performance despite the potential annotation bias. Second, a considerable portion of STS-Clinic sentence pairs are common descriptions that do not require comprehensive clinical knowledge to interpret the semantics. Typical examples include sentences extracted from Consultation Note or Discharge Summary as follows:

> *Plan: the patient stated an understanding of the program, and agrees to continue independently with a home management program.*
>
> *Thank you for choosing the name M.D. care team for your health care needs!*

On the other hand, there are many sentences in the STS-General associated with healthcare. An example is exhibited below:

> *Although obesity can increase the risk of health problems, skeptics argue, so do smoking and high cholesterol.*

Tang et al [30] have demonstrated that combining representations derived from different models is an efficient strategy in clinical STS. We explored similar strategies to combine sentence-level distributed representations, including vector concatenation, average pooling, max pooling, and convolution. Surprisingly, our results showed that such ensemble strategies did not help transformer-based STS systems. For example, for the ensemble model derived from the BERT-large and the XLNet-large models (ie, BERT-large + XLNet-large), the achieved Pearson correlation scores for vector concatenation, average pooling, max pooling, and convolution were 0.8764, 0.8760, 0.8799, and 0.8803, respectively. All the results were approximately 0.01 lower than that for XLNet-large (0.8864). We also observed that ensemble models' performances were consistently in between the two individual models (0.8549 for BERT-large and 0.8864 for XLNet-large). Future studies should examine this finding.

To examine the statistical significance among different models' results, we used a 1-tailed parametric test based on the Fisher Z-transformation [46], adopted in the previous SemEval STS shared tasks [2-4]. Our best model (ie, RoBERTa-large) achieved a statistically significant higher performance than most of our other solutions (see Multimedia Appendix 1) but was not significantly better than the models XLNet-large (P=.07), BERT-large + RoBERTa-large (P=.13), and RoBERTa-large + XLNet-large (P=.06). The significance analysis indicated that these four models performed very similarly to each other.

Error Analysis

We compared the system prediction from our best model (ie, RoBERTa-large) with the gold standards and identified sentence pairs with the largest discrepancy in terms of the similarity score. Among the top 50 sentence pairs, 26 of them had labeled scores in the range of 0.0 to 1.0, and only 6 sentence pairs had gold standard STS scores over 3.0. We further split the testing results into two subsets using a threshold score of 2.5 on gold standards and calculated the mean and median of the differences between the gold standards and predictions. For the subgroup consisting of sentence pairs with gold standard scores over 2.5, the mean and median of difference were 0.46 and 0.37. For the other subset (difference≤2.5), the mean and median of difference were 0.69 and 0.66. Therefore, it was more challenging for the system to predict appropriate STS scores for sentence pairs with low similarity (gold standard score≤2.5) than for those with high similarity.

We also observed that sentence pairs with high similarity scores usually have a similar sentence structure where many words occur in both sentences. Therefore, we hypothesized that the STS models will assign higher scores to sentence pairs that share a large portion of their lexicons and similar syntax. To test our hypothesis, we adopted the BertViz package [47] to profile the attention pattern of the RoBERTa-large model (ie, our best STS model). BertViz can generate the attention pattern between two sentences by linking words via lines, where the line weights reflect the attention weights; higher line weights indicate higher attention weights between the two words. Table 4 and Figure 3 show an example for two sentence pairs on a similar topic from the training and test sets. In the first example from the training set, the attention pattern has three dominant attention weights (eg, "questions-questions") and the similarity score for this sentence pair is labeled as 5.0. However, the attention pattern for the sentence pair from the test set also has similar dominant attention weights (such as "questions-questions") but was labeled with a similarity score of 0.0.

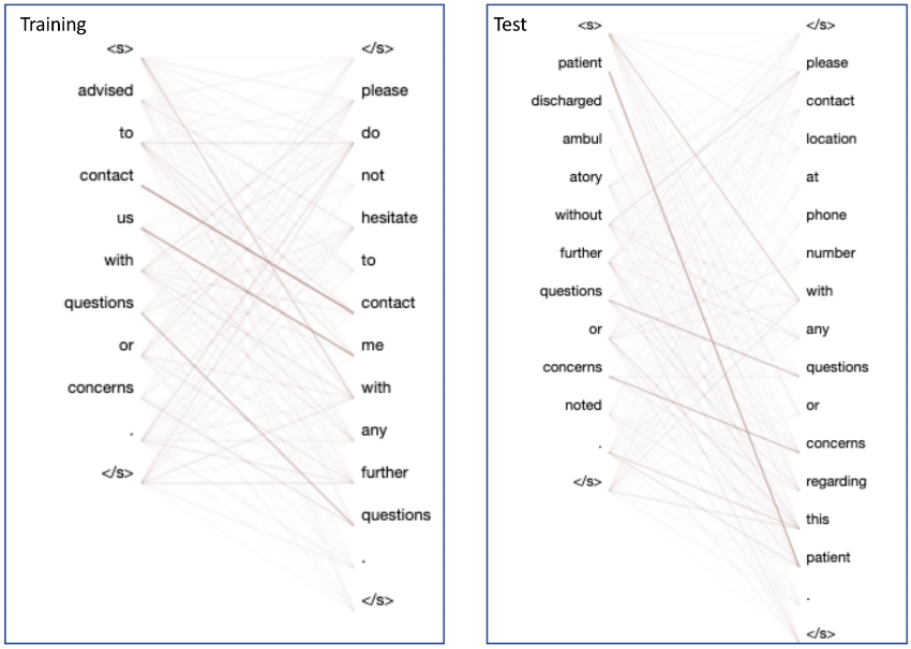
Table 4. Transformer model attention visualization on two examples from STS-Clinic.

Category	Sentence pair	Gold standard	Prediction
Training	• S1[a]: advised to contact us with questions or concerns. • S2: please do not hesitate to contact me with any further questions.	5	N/A[b]
Test	• S1: patient discharged ambulatory without further questions or concerns noted. • S2: please contact location at phone number with any questions or concerns regarding this patient.	0	2.5

[a]S: sentence.

[b]N/A: not applicable.

Figure 3. Transformer model attention visualization on two examples from STS-Clinic. STS: semantic textual similarity.



Limitations

This study has limitations. First, it is worth exploring methods to effectively integrate clinical resources with general English resources in transformer-based models. In this study, we explored an approach by pretraining transformer-based models with a clinical corpus (ie, MIMIC-III corpus). However, our results showed that this approach was not efficient. Therefore, new strategies to better integrate medical resources are needed. Second, our clinical STS systems performed better for sentence pairs with high similarity scores (ie, similarity score≥3 in gold standard) whereas, for the sentence pairs with low similarity scores (ie, similarity score<2 in gold standard), our systems still need to be improved. How to address this issue is one of our future focuses.

Conclusions

In this study, we demonstrated transformer-based models for measuring clinical STS and developed a system that can use various transformer algorithms. Our experiment results show that the RoBERTa model achieved the best performance compared to other transformer models. Our study demonstrated the efficiency of transformer-based models for assessing the semantic similarity for clinical text. Our models and system could be applied to various downstream clinical NLP applications. The source code, system, and pretrained models can be accessed on GitHub [48].

Acknowledgments

Research reported in this publication was supported by (1) the University of Florida Clinical and Translational Science Institute, which is supported in part by the National Institutes of Health (NIH) National Center for Advancing Translational Sciences under award number UL1TR001427; (2) the Patient-Centered Outcomes Research Institute under award number ME-2018C3-14754; (3) the Centers for Disease Control and Prevention under award number U18DP006512; (4) the NIH National Cancer Institute under award number R01CA246418; and (5) the NIH National Institute on Aging under award number R21AG061431-02S1.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and Patient-Centered Outcomes Research Institute.

We would like to thank the n2c2 organizers for providing the annotated corpus and the guidance for this challenge. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the graphics processing units used for this research.

Authors' Contributions

XY, JB, and YW were responsible for the overall design, development, and evaluation of this study. YM, HZ, and XH were involved in conducting experiments and result analysis. XY, JB, and YW wrote and edited this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 1-tailed parametric test results based on Fisher Z-transformation.
[XLSX File (Microsoft Excel File), 10 KB - medinform_v8i11e19735_app1.xlsx]

References

1. Agirre E, Cer D, Diab M, Gonzalez-Agirre A. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. : Association for Computational Linguistics; 2012 Presented at: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics; June 7-8, 2012; Montréal, Canada p. 385-393 URL: https://www.aclweb.org/anthology/S12-1051
2. Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. *SEM 2013 shared taskmantic Textual Similarity. : Association for Computational Linguistics; 2013 Presented at: *SEM 2013: The Second Joint Conference on Lexical and Computational Semantics; June 13-14, 2013; Atlanta, USA p. 32-43 URL: https://www.aclweb.org/anthology/S13-1004
3. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. 2014 Presented at: The 8th International Workshop on Semantic Evaluation (SemEval 2014); Aug 23-24, 2014; Dublin, Ireland p. 81-91 URL: https://www.aclweb.org/anthology/S14-2010 [doi: 10.3115/v1/s14-2010]
4. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. 2015 Presented at: The 9th International Workshop on Semantic Evaluation (SemEval 2015); June 4-5, 2015; Denver, USA p. 252-263 URL: https://www.aclweb.org/anthology/S15-2045/ [doi: 10.18653/v1/s15-2045]
5. Agirre E, Banea C, Cer D, Diab M, Gonzalez-Agirre A, Mihalcea R, et al. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. 2016 Presented at: The 10th International Workshop on Semantic Evaluation (SemEval-2016); June 16-17, 2016; San Diego, USA p. 497-511 URL: https://www.aclweb.org/anthology/S16-1081/ [doi: 10.18653/v1/S16-1081]
6. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. 2017 Presented at: The 11th International Workshop on Semantic Evaluation (SemEval-2017); Aug 3-4, 2017; Vancouver, Canada p. 1-14 URL: https://www.aclweb.org/anthology/S17-2001/ [doi: 10.18653/v1/s17-2001]
7. Farouk M. Measuring Sentences Similarity: A Survey. ArXiv 2019 Oct 06 [FREE Full text] [doi: 10.17485/ijst/2019/v12i25/143977]
8. Ramaprabha J, Das S, Mukerjee P. Survey on Sentence Similarity Evaluation using Deep Learning. In: J. Phys.: Conf. Ser. 2018 Apr 25 Presented at: National Conference on Mathematical Techniques and its Applications (NCMTA 18); Jan 5–6, 2018; Kattankulathur, India URL: https://iopscience.iop.org/article/10.1088/1742-6596/1000/1/012070 [doi: 10.1088/1742-6596/1000/1/012070]
9. Gomaa WH, Fahmy AA. A Survey of Text Similarity Approaches. IJCA 2013 Apr 18;68(13):13-18. [doi: 10.5120/11638-7118]
10. Béchara H, Costa H, Taslimipoor S, Gupta R, Orasan C, Corpas PG, et al. MiniExperts: An SVM Approach for Measuring Semantic Textual Similarity. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) Denver, Colorado: Association for Computational Linguistics; 2015 Presented at: The 9th International Workshop on Semantic Evaluation (SemEval 2015); Jun 4-5, 2015; Denver, USA p. 96-101 URL: https://www.aclweb.org/anthology/S15-2017/ [doi: 10.18653/v1/S15-2017]
11. Buscaldi D, Flores J, Ruiz I, Rodriguez I. SOPA: Random Forests Regression for the Semantic Textual Similarity task. 2015 Presented at: The 9th International Workshop on Semantic Evaluation (SemEval 2015); Jun 4-5, 2015; Denver, USA p. 132-137 URL: https://www.aclweb.org/anthology/S15-2024/ [doi: 10.18653/v1/s15-2024]

12. He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks. 2015 Presented at: The 2015 Conference on Empirical Methods in Natural Language Processing; Sept 19 – 21, 2015; Lisbon, Portugal p. 1576-1586 URL: https://www.aclweb.org/anthology/D15-1181/ [doi: 10.18653/v1/d15-1181]
13. Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. : AAAI Press; 2016 Presented at: The Thirtieth AAAI Conference on Artificial Intelligence; Feb 12–17, 2016; Phoenix, USA p. 2786-2792 URL: https://dl.acm.org/doi/10.5555/3016100.3016291 [doi: 10.5555/3016100.3016291]
14. Kashyap A, Han L, Yus R, Sleeman J, Satyapanich T, Gandhi S, et al. Robust semantic text similarity using LSA, machine learning, and linguistic resources. Lang Resources & Evaluation 2015 Oct 30;50(1):125-161. [doi: 10.1007/s10579-015-9319-2]
15. Niraula N, Banjade R, Ştefănescu D, Rus V. Experiments with Semantic Similarity Measures Based on LDA and LSA. 2013 Presented at: The First International Conference on Statistical Language and Speech Processing (SLSP 2013); July 29-31, 2013; Tarragona, Spain p. 188-199 URL: https://link.springer.com/chapter/10.1007/978-3-642-39593-2_17 [doi: 10.1007/978-3-642-39593-2_17]
16. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018 [FREE Full text]
17. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv 2019 [FREE Full text]
18. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv 2019 [FREE Full text]
19. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv 2019 [FREE Full text]
20. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv 2019 [FREE Full text]
21. Wang W, Bi B, Yan M, Wu C, Bao Z, Xia J, et al. StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. arXiv 2019 [FREE Full text]
22. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 2015 Mar 18;36:345-359. [doi: 10.1146/annurev-publhealth-031914-122747] [Medline: 25581157]
23. Obeid JS, Beskow LM, Rape M, Gouripeddi R, Black RA, Cimino JJ, et al. A survey of practices for the use of electronic health records to support research recruitment. J Clin Transl Sci 2017 Aug;1(4):246-252 [FREE Full text] [doi: 10.1017/cts.2017.301] [Medline: 29657859]
24. 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]
25. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
26. Zhang R, Pakhomov S, McInnes BT, Melton GB. Evaluating measures of redundancy in clinical texts. AMIA Annu Symp Proc 2011;2011:1612-1620 [FREE Full text] [Medline: 22195227]
27. Wang MD, Khanna R, Najafi N. Characterizing the Source of Text in Electronic Health Record Progress Notes. JAMA Intern Med 2017 Aug 01;177(8):1212-1213 [FREE Full text] [doi: 10.1001/jamainternmed.2017.1548] [Medline: 28558106]
28. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Jt Summits Transl Sci Proc 2010 Mar 01;2010:1-5 [FREE Full text] [Medline: 21347133]
29. Soganciogl G, Öztürk H, Özgür A. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics 2017 Jul 15;33(14):i49-i58 [FREE Full text] [doi: 10.1093/bioinformatics/btx238] [Medline: 28881973]
30. Xiong Y, Chen S, Qin H, Cao H, Shen Y, Wang X, et al. Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. BMC Med Inform Decis Mak 2020 Apr 30;20(Suppl 1):72 [FREE Full text] [doi: 10.1186/s12911-020-1045-z] [Medline: 32349764]
31. Chen Q, Du J, Kim S, Wilbur WJ, Lu Z. Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. BMC Med Inform Decis Mak 2020 Apr 30;20(Suppl 1):73 [FREE Full text] [doi: 10.1186/s12911-020-1044-0] [Medline: 32349758]
32. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resources & Evaluation 2018 Oct 24;54(1):57-72. [doi: 10.1007/s10579-018-9431-1]
33. Rastegar-Mojarad M, Liu S, Wang Y, Afzal N, Wang L, Shen F, et al. BioCreative/OHNLP Challenge 2018. 2018 Presented at: The 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB); Aug 29 - Sept 1, 2018; Washington DC, USA p. 575 URL: https://doi.org/10.1145/3233547.3233672 [doi: 10.1145/3233547.3233672]
34. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. Overview of the 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity. JMIR Medical Informatics 2020 [FREE Full text] [doi: 10.2196/23375]

35. Zhang C, Ma Y, editors. Ensemble Learning. In: Ensemble machine learning: methods and applications. New York, USA: Springer; 2012.
36. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv 2019 [FREE Full text]
37. 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;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
38. 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;17(3):261-272 [FREE Full text] [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
39. Murphy K. Machine Learning: A Probabilistic Perspective. Cambridge, USA: MIT Press; 2012.
40. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
41. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv 2019 [FREE Full text]
42. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. arXiv 2019 [FREE Full text]
43. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. 2019 Presented at: The 18th BioNLP Workshop and Shared Task (BioBLP 2019); Aug 1, 2019; Florence, Italy p. 58-65. [doi: 10.18653/v1/w19-5006]
44. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019 Nov 01;26(11):1297-1304. [doi: 10.1093/jamia/ocz096] [Medline: 31265066]
45. Tawfik NS, Spruit MR. Evaluating sentence representations for biomedical text: Methods and experimental results. J Biomed Inform 2020 Apr;104:103396. [doi: 10.1016/j.jbi.2020.103396] [Medline: 32147441]
46. Fisher RA. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika 1915 May;10(4):507. [doi: 10.2307/2331838]
47. Vig J. A Multiscale Visualization of Attention in the Transformer Model. arXiv 2019 Jun 12 [FREE Full text] [doi: 10.18653/v1/p19-3007]
48. 2019 N2C2 Track-1 Clinical Semantic Textual Similarity. GitHub. URL: https://github.com/uf-hobi-informatics-lab/2019_N2C2_Track1_ClinicalSTS.git [accessed 2020-11-02]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
MIMIC-III: Medical Information Mart for Intensive Care
MLM: masked language modeling
n2c2: National Natural Language Processing Clinical Challenges
NIH: National Institutes of Health
NLP: natural language processing
OHNLP: Open Health Natural Language Processing
RoBERTa: Robustly optimized BERT approach
SemEval: semantic evaluation
STS: semantic textual similarity

Edited by Y Wang; submitted 27.07.20; peer-reviewed by F Li, J Lei, A Mavragani; comments to author 06.10.20; revised version received 19.10.20; accepted 26.10.20; published 23.11.20.

Please cite as:
Yang X, He X, Zhang H, Ma Y, Bian J, Wu Y
Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models
JMIR Med Inform 2020;8(11):e19735
URL: http://medinform.jmir.org/2020/11/e19735/
doi:10.2196/19735
PMID:33226350

Original Paper

Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning

Diwakar Mahajan[1], MS; Ananya Poddar[1], MS; Jennifer J Liang[1], MD; Yen-Ting Lin[2], BS; John M Prager[3], PhD; Parthasarathy Suryanarayanan[1], BTECH; Preethi Raghavan[1], PhD; Ching-Huei Tsou[1], PhD

[1]IBM Research, Yorktown Heights, NY, United States

[2]National Taiwan University, Taipei, Taiwan

[3]Formerly IBM Research, Yorktown Heights, NY, United States

Corresponding Author:
Diwakar Mahajan, MS
IBM Research
1101 Kitchawan Road
Yorktown Heights, NY, 10598
United States
Phone: 1 914 945 1614
Email: dmahaja@us.ibm.com

Abstract

Background: Although electronic health records (EHRs) have been widely adopted in health care, effective use of EHR data is often limited because of redundant information in clinical notes introduced by the use of templates and copy-paste during note generation. Thus, it is imperative to develop solutions that can condense information while retaining its value. A step in this direction is measuring the semantic similarity between clinical text snippets. To address this problem, we participated in the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing Consortium (OHNLP) clinical semantic textual similarity (ClinicalSTS) shared task.

Objective: This study aims to improve the performance and robustness of semantic textual similarity in the clinical domain by leveraging manually labeled data from related tasks and contextualized embeddings from pretrained transformer-based language models.

Methods: The ClinicalSTS data set consists of 1642 pairs of deidentified clinical text snippets annotated in a continuous scale of 0-5, indicating degrees of semantic similarity. We developed an iterative intermediate training approach using multi-task learning (IIT-MTL), a multi-task training approach that employs iterative data set selection. We applied this process to bidirectional encoder representations from transformers on clinical text mining (ClinicalBERT), a pretrained domain-specific transformer-based language model, and fine-tuned the resulting model on the target ClinicalSTS task. We incrementally ensembled the output from applying IIT-MTL on ClinicalBERT with the output of other language models (bidirectional encoder representations from transformers for biomedical text mining [BioBERT], multi-task deep neural networks [MT-DNN], and robustly optimized BERT approach [RoBERTa]) and handcrafted features using regression-based learning algorithms. On the basis of these experiments, we adopted the top-performing configurations as our official submissions.

Results: Our system ranked first out of 87 submitted systems in the 2019 n2c2/OHNLP ClinicalSTS challenge, achieving state-of-the-art results with a Pearson correlation coefficient of 0.9010. This winning system was an ensembled model leveraging the output of IIT-MTL on ClinicalBERT with BioBERT, MT-DNN, and handcrafted medication features.

Conclusions: This study demonstrates that IIT-MTL is an effective way to leverage annotated data from related tasks to improve performance on a target task with a limited data set. This contribution opens new avenues of exploration for optimized data set selection to generate more robust and universal contextual representations of text in the clinical domain.

(JMIR Med Inform 2020;8(11):e22508) doi:10.2196/22508

KEYWORDS

electronic health records; semantic textual similarity; natural language processing; multi-task learning; transfer learning; deep learning

Introduction

Background

The wide adoption of electronic health records (EHRs) has led to clinical benefits with increased efficiency and financial benefits [1]. Although electronic documentation has greatly improved the legibility and accessibility of clinical documentation, the use of templates and copy-paste during note generation has inadvertently introduced unnecessary, redundant, and potentially erroneous information (ie, note bloat), resulting in decreased readability and functional usability of the generated clinical notes [2-5]. A previous study [6] on 23,630 clinical notes identified that in a typical note, only 18% of the text was manually entered, whereas 46% was copied and 36% imported. This problem of note bloat not only increases physician cognitive burden [7] but also becomes a challenge for the secondary use of EHRs in clinical informatics [8]. Figure 1 illustrates this challenge with an example of 2 sample clinical notes from the same patient from consecutive visits; blue and yellow highlighted text indicate content that have been added or modified, respectively, whereas the plain unhighlighted text indicates information that is the same across clinical notes.

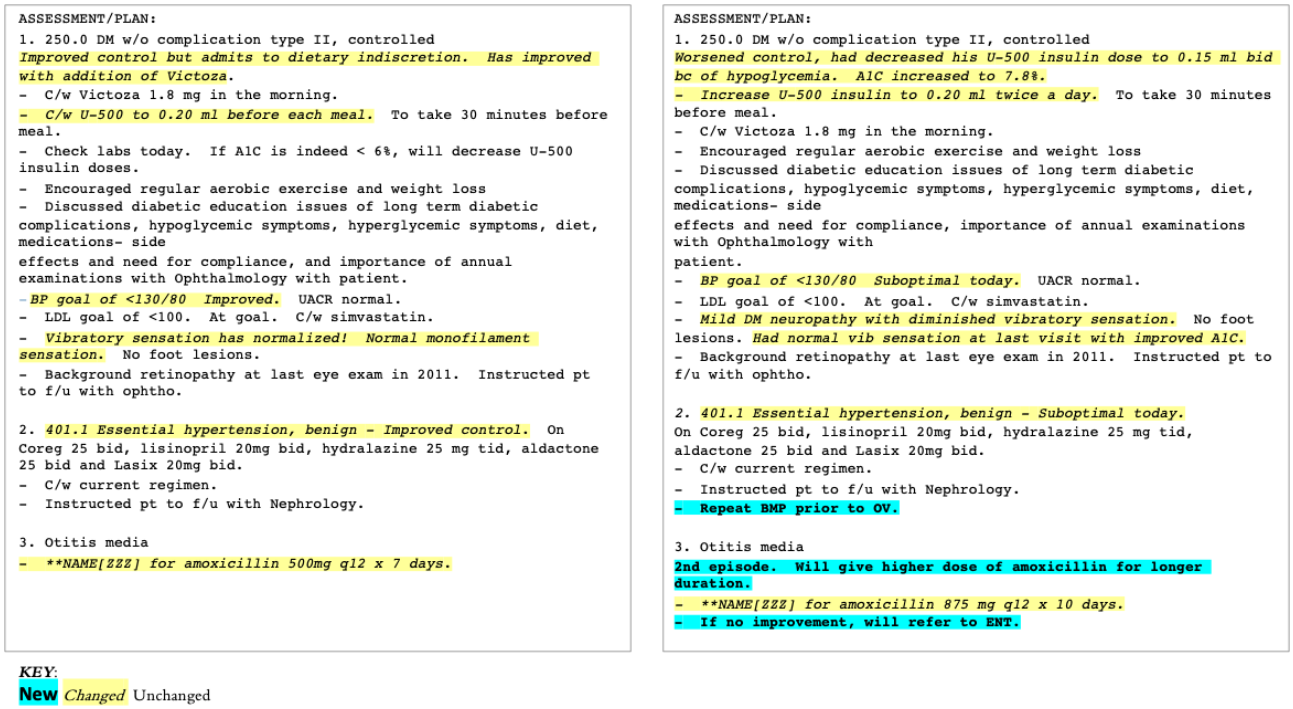
One way to minimize data redundancy and highlight new information in unstructured clinical notes can be to compute the semantic similarity between clinical text snippets. This process of measuring the degree of semantic equivalence between clinical text snippets is known as clinical semantic textual similarity [9]. As semantic textual similarity (STS) is a foundational language understanding problem, successful modeling of this task may help improve other higher-level applications in the clinical domain [9], such as clinical question answering with evidence-based retrieval, clinical text summarization, semantic search, conversational systems, and clinical decision support.

The 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing Consortium (OHNLP) track on clinical semantic textual similarity (ClinicalSTS) [10] was organized to tackle this specific task: given a pair of clinical text snippets, assign a numerical score from 0 to 5 to indicate the degree of semantic similarity. This is an extension of a previous challenge from BioCreative/OHNLP 2018 ClinicalSTS [11,12] that was inspired by the Semantic Evaluation (SemEval) semantic textual similarity (STS) shared tasks [13-18], which have been organized since 2012 in the general domain.

Pretrained language models have been shown to be effective for achieving state-of-the-art results on many general and clinical domain natural language processing (NLP) tasks [19], including STS. However, when the target domain differs substantially from the pretraining corpus, the contextualized embeddings may be ineffective for the target task. Furthermore, when the amount of training data are limited, as is common for clinical NLP tasks, fine-tuning experiments are potentially brittle and rely on the pretrained encoder parameters to be reasonably close to an ideal setting for the target task [20]. A previous study has shown that small training data sets can significantly benefit from an intermediate training step [20]. In a complementary work, multi-task learning (MTL) [21] has been shown to be effective in leveraging supervised data from multiple related tasks for a target task. Furthermore, it has been observed that MTL and language model pretraining are complementary technologies [21].

On the basis of these observations, we present a novel methodology that iteratively performs intermediate training of a pretrained language model in an MTL setup using related data-rich tasks. In this iterative process, related data sets were purposefully selected to induce representative knowledge of the target task. In addition, we evaluated the impact of combining multiple transformer-based language models pretrained on diverse corpora. Our system ranked first in the 2019 n2c2/OHNLP ClinicalSTS challenge, achieving state-of-the-art results.

Figure 1. Two sample clinical notes for the same patient from consecutive visits. Plain text indicates same content between 2 notes; italics (yellow highlight) indicate the content that has been modified, and bold (blue highlight) indicates new content in the second note.



Relevant Literature

STS is defined as the comparison of a pair of text snippets, approximately one sentence in length, resulting in a numerical score that takes a value on a continuous scale of 0 to 5, indicating degrees of semantic similarity [9,18]. STS, along with paraphrase detection and textual entailment, is a form of semantic relatedness task. Paraphrase detection is the identification of sentences that are semantically identical [22], whereas textual entailment is the task of reasoning if one text snippet can be inferred from another [23-25]. STS is more similar to paraphrase detection because of the symmetricity of the relationship, as compared with entailment, which is asymmetric. However, unlike paraphrase detection, STS expands on the binary output scoring in paraphrase detection to capture gradations of relatedness.

Early research on STS, in both the general and clinical domains, focused on lexical semantics, basic syntactic similarity, surface form matching, and alignment-based methods [26-28]. The overarching theme behind these methods is the identification, alignment, and scoring of semantically related words and phrases and aggregating their scores. However, the absence of a principled way of combining the topological and semantic information led to the construction of sentence representations by building a linear composition of the distributed representations of individual words [29-32]. Although these techniques were an improvement over traditional approaches, they fell short as they did not take the surrounding context into account while generating distributed representations.

Early attempts at building richer representations that encode several linguistic aspects of a sentence for computing similarity included paragraph vectors [33-36], word embedding weighting and principal component removal [37], and convolutional deep structured semantic model [38,39]. However, recent studies on pretrained language models have achieved a breakthrough in sentence representation learning [19,40,41]. Bidirectional encoder representations from transformers (BERT) build upon the ideas from the transformer [42] to construct rich sentence representations and has achieved state-of-the-art results on many general and clinical domain NLP tasks [24,43]. In this process, a transformer-based model is first pretrained on large corpora to learn universal language representations and is then fine-tuned with a task-specific output layer for the target task. BERT has been adapted to biomedical (bidirectional encoder representations from transformers for biomedical text mining [BioBERT]) [44] and clinical (bidirectional encoder representations from transformers on clinical text mining [ClinicalBERT]) domains [45,46].

The performance of BERT and its domain-specific variants could be further improved through MTL. MTL [47] refers to training a model simultaneously for multiple related tasks, and MTL benefits from a regularization effect by alleviating overfitting to a specific task, thus making the learned representations universal across tasks. Supplementary training on intermediate tasks refers to the second stage of pretraining of a model, with data-rich intermediate supervised tasks. Recent studies, such as multi-task deep neural networks (MT-DNN) [21] and supplementary training on intermediate labeled-data tasks [20], show that the use of MTL and intermediate pretraining generates more robust and universal learned representations, resulting in better domain adaptation with fewer in-domain labels.

The winning systems in ClinicalSTS 2018 challenge [48] and SemEval 2017 [49] built upon a combination of approaches referenced earlier in this section. In general, they employed ensembled feature engineering methods (random forest, gradient

boosting, and XGBoost) with features based on n-gram overlap, edit distance, longest common prefix/suffix/substring, word alignments [50,51], summarization and machine translation evaluation metrics, and deep learning [36,52]. In contrast to these systems, our study builds upon the modern neural approaches referenced earlier. Specifically, our system implements MTL and supplementary training on intermediate labeled tasks with ClinicalBERT to achieve state-of-the-art performance on the ClinicalSTS 2019 task. Following the demonstration of our system at the 2019 n2c2/OHNLP challenge presentation, additional systems leveraging MTL in ClinicalBERT [53,54] have been implemented with promising results.

Methods

Data Set

The 2019 ClinicalSTS data set was prepared by the n2c2/OHNLP challenge organizers from sentences collected from clinical notes in the Mayo Clinic's clinical data warehouse. Candidate sentence pairs were then generated using an average value ≥0.45 of surface lexical similarity methods, namely, Ratcliff/Obershelp [55], cosine similarity, and Levenshtein distance. This resulted in 2054 pairs, of which 1642 were released as the training set and the remaining 412 were held by the organizers for testing. Protected health information was removed using a mix of frequency filtering approach [56] and manual review process. Each sentence pair was independently reviewed by 2 clinical experts and scored on a scale of 0 to 5 based on their semantic equivalence (0 for no semantic equivalence to 5 for complete semantic equivalence). Interannotator agreement was 0.6 based on weighted Cohen kappa. The averaged score between the 2 annotators was used as the gold standard. Table 1 presents a few examples from the data set.

We split the provided training data set of 1642 sentence pairs into 75.03% (1232/1642), 14.98% (246/1642), and 9.99% (164/1642) to form our train, validation, and internal test data sets, respectively.

Table 1. Sample sentence pairs and annotations from the clinical semantic textual similarity data set.

Ground truth[a]		Score	Observations	
Sentence 1	Sentence 2		Domain dependence	Comments
"The patient was taken to the *PACU*[b] in a stable condition."	"The patient was taken to the *post anesthesia care unit* postoperatively for recovery."	5.0	Domain specific	Clinical abbreviations
"*Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA*[c]*Aerosol 1-2 puffs by inhalation every 4 hours as needed*."	"*Ipratropium-Albuterol [COMBIVENT] 18-103 mcg/Actuation Aerosol 2 puffs by inhalation two times a day as needed*"	3.5	Domain specific	Medication instruction parsing
"Cardiovascular assessment findings include *heart rate normal, atrial fibrillation with controlled ventricular response*."	"Cardiovascular assessment findings include *heart rate, first degree AV*[d]*Block*."	3.0	Domain specific	Medical concept similarity and medical concept mapping
"He was *prepped and draped in the standard* fashion."	"The affected shoulder was *prepared and draped with the usual* sterile technique."	3.0	Domain independent	Alignment
"Musculoskeletal: *Positive* for gait problem, joint swelling and extremity pain."	"Musculoskeletal: *Negative* for back pain, myalgias and extremity pain."	1.5	Domain independent	Assertion classification (polarity)

[a]Italics indicate the phrases within each sentence which correspond to the observations.

[b]PACU: post anesthesia care unit.

[c]HFA: hydrofluoroalkane.

[d]AV: atrioventricular.

Analysis of this data set revealed 2 characteristics that we consider in our approach to this task. First, the lack of sufficient training data makes it difficult to train robust machine learning models using only the given training data. Second, clinical semantic similarity relies on both domain-specific (eg, clinical abbreviation expansion, medical concept detection, and medical concept normalization) and domain-independent (eg, assertion classification and alignment detection) aspects, as demonstrated by the sample sentence pairs in Table 1. For the first sentence pair, a domain-specific understanding of PACU as an abbreviation for post anesthesia care unit is necessary to infer the high semantic equivalence. For the fourth sample sentence pair, domain-independent understanding of the difference in polarity between Positive and Negative is necessary to infer the low similarity equivalence.

To address the lack of sufficient training data and leverage the domain-specific and domain-independent aspects of clinical semantic similarity, we propose an approach that combines the following:

- an iterative intermediate multi-task training step for effective transfer learning employing other related annotated data sets
- an ensemble module that combines language models pretrained on both domain-specific and domain-independent data sets and also incorporates other features.

Iterative Intermediate Training Using MTL

We performed iterative multi-task training on a transformer-based language model using annotated data sets from related tasks to induce representative knowledge of the

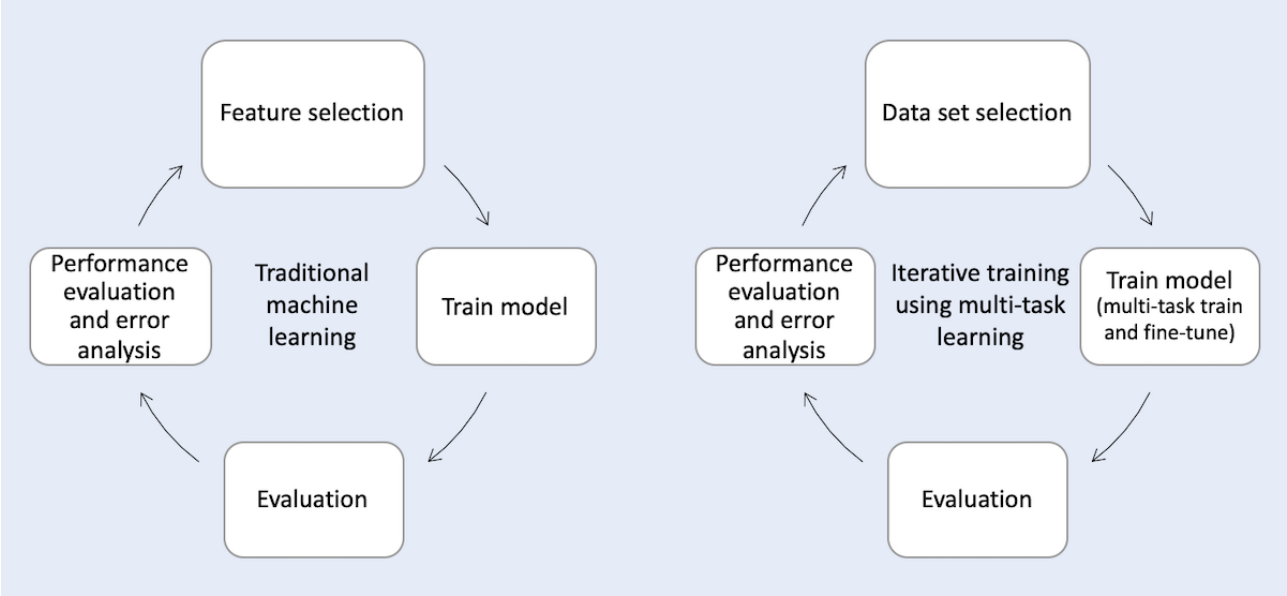
target task. With each iteration, annotated data sets from related tasks were added or removed. Following data set selection, the language model was then trained using MTL on the selected data sets, fine-tuned on the target task, and its results were evaluated and error analysis was performed to determine the data set selection for the next iteration. We refer to this entire process as iterative intermediate training using multi-task learning (IIT-MTL).

IIT-MTL is analogous to traditional feature-based machine learning methodologies, where performance evaluation and error analysis lead to feature selection used to train the model. In IIT-MTL, instead of feature selection, data set selection is employed to select data sets. Figure 2 presents IIT-MTL compared with the traditional machine learning approach.

For the ClinicalSTS task, ClinicalBERT was used as our base model as it was pretrained on a clinical corpus and provides clinically specific contextual embeddings most suited to our task. Through IIT-MTL, a refined clinical domain-specific language model, IIT-MTL on ClinicalBERT (IIT-MTL-ClinicalBERT), is obtained that has been iteratively tuned for high performance on the ClinicalSTS task.

In the following sections, we present each step of IIT-MTL as applied to the ClinicalSTS task: (1) the data set selection process, including details of each iteration and data sets used; (2) the MTL architecture with the task-specific layers considered during the iterative process; and (3) fine-tuning on the target task.

Figure 2. Comparison of traditional machine learning approach (left), where performance evaluation and error analysis lead to feature selection, and our proposed iterative training using multi-task learning approach (right), where performance evaluation and error analysis lead to data set selection.



Data Set Selection

For effective performance on the target ClinicalSTS task, we not only trained our model using MTL as an intermediate step but also iteratively selected the data sets employed during this process based on error analysis of the performance on the target task. The selection of complementary data sets is critical to this process as it significantly impacts the contextual representations in the final model.

Several publicly available data sets were considered in these iterations, including Semantic Textual Similarity Benchmark (STS-B) [18], Recognizing Question Entailment (RQE) [57], natural language inference data set for the clinical domain (MedNLI) [24], and Quora Question Pairs (QQP) [58]. STS-B consists of 8.6 K sentence pairs drawn from news headlines, video and image captions, and natural language inference data, each annotated with a score of 0 to 5 to indicate the degree of semantic equivalence. RQE consists of 8.9 K pairs of clinical questions, each annotated with a binary value to indicate entailment (or lack of) between the 2 questions. MedNLI consists of 14 K sentences extracted from clinical notes in the Medical Information Mart for Intensive Care (MIMIC-III) database [59], with each sentence pair annotated as either entailment, neutral, or contradiction. QQP consists of 400 K pairs of questions extracted from the Quora question-and-answer website, each annotated with a binary value to indicate the similarity (or lack of) between the 2 questions. We created 2 additional data sets for use in IIT-MTL for ClinicalSTS: a sentence topic-based data set (Topic) and a medication named entity recognition data set (MedNER). Topic was created on sentences within the ClinicalSTS data set, where each sentence was manually annotated with a label from a predefined list of topics (eg, MED, SIGNORSYMPTOM, EXPLAIN, and OTHER). MedNER was autogenerated using a medication extraction tool [60] on 1000 randomly selected clinical notes in the MIMIC-III database to recognize medications and its related artifacts (eg, strength, form, frequency, route, dosage, and duration). A summary of all data sets used is presented in Table 2, with additional details provided in Multimedia Appendix 1 [10,18,24,57,59-62].

Table 2. Data sets used in multi-task learning.

Data set	Task	Domain	Size	Example
STS-B[a]	Sentence pair similarity	General	8600	Sentence 1: "A young child is riding a horse"; Sentence 2: "A child is riding a horse"; Similarity: 4.75
RQE[b]	Sentence pair classification	Biomedical	8900	Sentence 1: "Doctor X thinks he is probably just a normal 18 month old but would like to know if there are a certain number of respiratory infections that are considered normal for that age"; Sentence 2: "Probably a normal 18 month old but how many respiratory infections are normal"; Ground truth: entailment
MedNLI[c]	Sentence pair classification	Clinical	14,000	Sentence 1: "Labs were notable for Cr 1.7 (baseline 0.5 per old records) and lactate 2.4"; Sentence 2: "Patient has normal Cr"; Ground truth: contradiction
QQP[d]	Sentence pair classification	General	400,000	Sentence 1: "Why do rockets look white?"; Sentence 2: "Why are rockets and boosters painted white?"; Ground truth: 1
Topic	Sentence classification	Clinical	1,300,000	Sentence: "Negative for difficulty urinating, pain with urination, and frequent urination"; Ground truth: SIGNORSYMPTOM
MedNER[e]	Token-wise classification	Clinical	15,000	Sentence: "he developed respiratory distress on the AM[f] of admission, cough day PTA[g], CXR[h] with B/L[i] LL[j] PNA[k], started ciprofloxacin and levofloxacin"; Ground truth: ciprofloxacin [DRUG] levofloxacin [DRUG]

[a]STS-B: semantic textual similarity benchmark.

[b]RQE: Recognizing Question Entailment.

[c]MedNLI: natural language inference data set for the clinical domain.

[d]QQP: Quora Question Pairs.

[e]MedNER: medication named entity recognition.

[f]AM: morning.

[g]PTA: prior to admission.

[h]CXR: chest x-ray.

[i]B/L: bilateral.

[j]LL: left lower.

[k]PNA: pneumonia.

We established 2 baselines by fine-tuning 2 pretrained language models, BERT and ClinicalBERT, on the target ClinicalSTS task. Using the stronger baseline of ClinicalBERT, a total of 5 iterations were performed in IIT-MTL for the ClinicalSTS task. The selection of data sets for each iteration was decided based on our understanding of the ClinicalSTS task and error analysis of the results of the previous iteration. The data set selection for each iteration is detailed as follows. For each iteration, D indicates the set of data sets used for multi-task training, following which the model is further fine-tuned to the target ClinicalSTS task and evaluated before the next iteration.

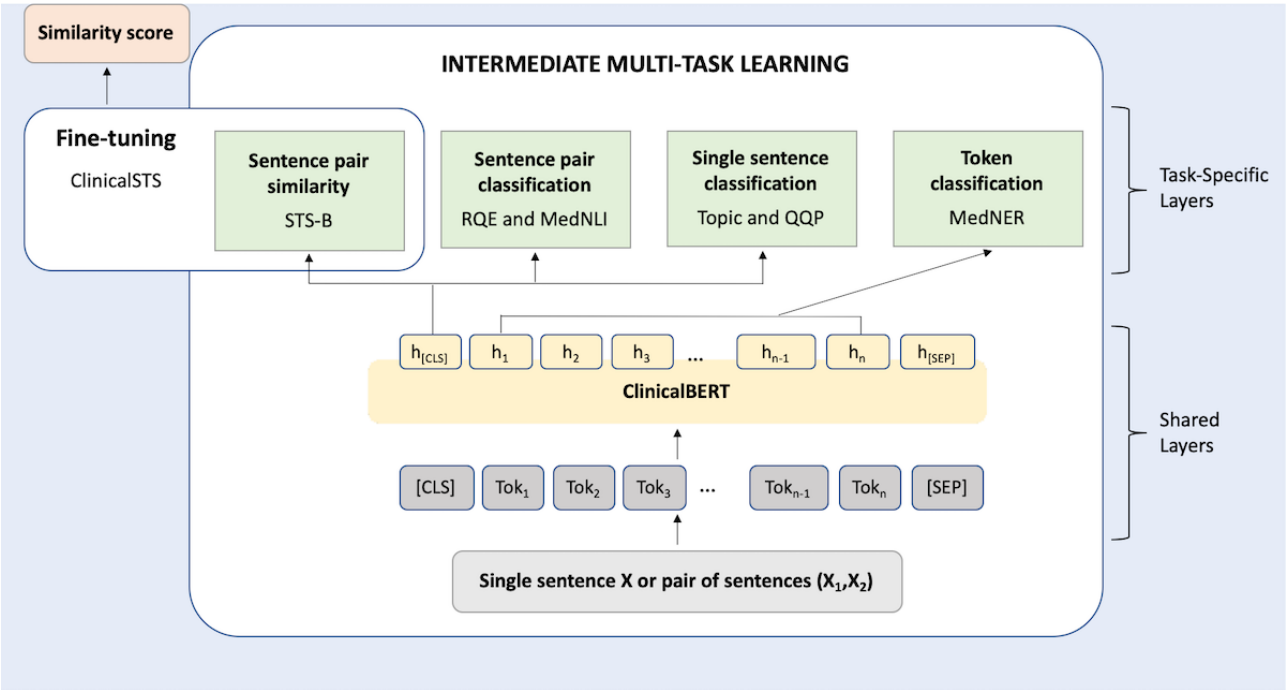
- *Iteration 1: D={STS-B}*: STS-B was employed for multi-task training because it conforms to the same task (STS) in the general domain.
- *Iteration 2: D={STS-B, RQE, MedNLI}*: Next, we added RQE and MedNLI, which are sentence pair classification tasks in the clinical domain, and, hence, are similar to our target task from a domain perspective.
- *Iteration 3: D={STS-B, RQE, MedNLI, Topic}*: Analysis of the output from iteration 2 showed that sentence pairs on different topics within ClinicalSTS express similarity in different ways. Thus, we created and added the Topic data set.
- *Iteration 4: D={STS-B, RQE, MedNLI, Topic, MedNER}*: Analysis of the output from iteration 3 showed that medication instruction sentences (eg, "Tylenol tablet 2 tablets by mouth as needed.") were the worst performing sentence pairs. To induce medication-related knowledge, we created and added the MedNER data set to the mix.
- *Iteration 5*: D=*{STS-B, RQE, MedNLI, Topic, MedNER, QQP}*: QQP was added in our final iteration as it is a sentence pair classification task, although in the general domain.

The final set of data sets used in the model for the ClinicalSTS task (IIT-MTL-ClinicalBERT) was determined based on the performance analysis of each iteration.

Intermediate MTL Architecture

The architecture of our intermediate MTL setup is shown in Figure 3 and is based on the process specified in the study by Liu et al [21].

Figure 3. Intermediate multi-task learning and fine-tuning architecture. ClinicalSTS: clinical semantic textual similarity; STS-B: semantic textual similarity benchmark; RQE: recognizing question entailment; MedNLI: natural language inference data set for the clinical domain; QQP: Quora question pairs; MedNER: medication named entity recognition data set; ClinicalBERT: bidirectional encoder representations from transformers on clinical text mining.



The lower shared layers are based on BERT-base architecture [19], whereas the higher segregated layers represent task-specific outputs. The task-specific layers correspond to the data sets selected during the data set selection.

The input can either be a single sentence (X) or a pair of sentences (X_1, X_2) delimited with the separating token ([SEP]). All input texts are tokenized using WordPieces [63] and truncated to spans no longer than 512 tokens. Following this, tokens are added to the start ([CLS]) and end ([SEP]) of the input. In the shared layers, a lexicon encoder converts the input into a sequence of input embedding vectors, one for each token. Next, a transformer encoder captures the contextual information and generates a sequence of contextual embeddings. This semantic representation is shared across all tasks and feeds into multiple lightweight task-specific architectures, each implementing a different task objective. In the training phase, we fine-tuned the shared layers along with task-specific layers using the multi-task objectives, detailed below:

- *Sentence Pair Similarity*: Suppose $h_{[CLS]}$ is the contextual embedding of [CLS] for input sentence pair (X_1, X_2) and w_{SPS} is a task-specific parameter vector. We utilized a fully connected layer to compute the similarity score , where  is a real value of range (−∞, ∞). We use the mean squared error as the objective function:

where y is the similarity score for the sentence pair.

- *Single Sentence Classification*: Suppose $h_{[CLS]}$ is the contextual embedding of [CLS] for input sentence X and w_{SSC} is a task-specific parameter vector. The probability that X is labeled as class c is predicted by logistic regression with softmax:

This task is trained using the cross-entropy loss as the objective:

where  is the binary indicator (0 or 1) if the class label c is the correct classification for X.

- *Sentence Pair Classification*: Suppose $h_{[CLS]}$ is the contextual embedding of [CLS] for sentence pair (X_1, X_2) and w_{SPC} is a task-specific parameter vector. As the two sentences are packed together, we can predict that the relation R between X_1 and X_2 is given as  similar to single sentence classification. We trained the task using the cross-entropy loss as specified previously
- *Token Classification*: Suppose $h_{[1:n]}$ is the contextual embedding for tokens $Tok_{[1:n]}$ in packed sentence pair (X_1, X_2) and w_{TC} is a task-specific parameter vector. The token classification is trained using a per-entity linear classifier, where the probability that $Tok_{[j]}$ labeled as class c is predicted by logistic regression with softmax:  . Here,  . This task is trained using the cross-entropy loss as specified previously.

The process for training our intermediate MTL architecture is demonstrated in Textbox 1. We initialized the shared layers of

our architecture with the parameters of the pretrained ClinicalBERT [46]. The task-specific layers were randomly initialized. We jointly refer to them as θ. Next, we created equal-sized subsamples (mini-batches) from each data set. For every epoch, a mini-batch b_t was selected (from each of the MTL data sets detailed previously), and the model was updated according to the task-specific objective for task t. We used the mini-batch–based stochastic gradient descent to update the parameters. A detailed explanation of the training parameters is provided in Multimedia Appendix 2 [19,21,63-65].

Textbox 1. Multi-task learning algorithm.

```
Initialize model parameters θ
Create E by merging mini-batches (b_t) for each data set in D
for epoch in 1,2,….., epoch_max do
    Shuffle E
    for b_t in E do
        Compute loss: L (θ) based on task t;
        Compute gradient: ∇(θ)
        Update model: θ=θ−η∇(θ)
    end
end
```

Fine-Tuning

After multi-task training, we fine-tuned the model on the target ClinicalSTS task. As ClinicalSTS is a sentence similarity task, we fine-tuned the sentence pair similarity task-specific layer of the multi-task architecture (Figure 3) to train the model using the ClinicalSTS data set. The predictions on the internal test data set were evaluated, which drove the data set selection process. A detailed explanation of the training parameters is provided in Multimedia Appendix 2.

Ensemble Module

To induce both domain-specific and domain-independent aspects of clinical semantic similarity, we leveraged other pretrained language models in addition to IIT-MTL-ClinicalBERT in the ensemble module. During this process, we fine-tuned other pretrained language models on the target task, ensembled their predictions with predictions from IIT-MTL-ClinicalBERT (which was already fine-tuned during IIT-MTL), and then incorporated additional similarity features. In the following sections, we describe the (1) language models used, (2) additional similarity features incorporated, and (3) different ensembling techniques explored.

Language Models

A total of 4 language models were used in our ensemble module: IIT-MTL-ClinicalBERT, BioBERT [44], MT-DNN [21], and robustly optimized BERT approach (RoBERTa) [66]. IIT-MTL-ClinicalBERT, the output of IIT-MTL, was derived from ClinicalBERT [46], and therefore, it provided clinical domain-specific contextual embeddings. To provide contextual representations from a similar but slightly different domain, we used BioBERT, which is also BERT-based but has been further pretrained on the biomedical corpus. To account for the domain-independent aspects of clinical semantic similarity, we used language models from the general domain, specifically RoBERTa and MT-DNN. RoBERTa is based on BERT but has been optimized for better performance, whereas MT-DNN leverages large amounts of cross-task data, resulting in more generalized and robust text representations. We selected RoBERTa and MT-DNN for use in our ensemble module because at the time of the 2019 n2c2/OHNLP challenge, they achieved state-of-the-art results on multiple tasks similar to ClinicalSTS, including STS-B [43], Multi-Genre Natural Language Inference [23], Question answering Natural Language Inferencing [67], and Recognizing Textual Entailment [68]. Table 3 presents an overview of the language models used in our experiments.

Table 3. Pretrained language models used in the ensemble module and their training corpora.

Language model	Corpora for language model pretraining	Domain
MT-DNN[a]	Wikipedia+BookCorpus	General
RoBERTa[b]	Wikipedia+BookCorpus+CC-News+OpenWebText+Stories	General
BioBERT[c]	Wikipedia+BookCorpus+PubMed+PMC[d]	Biomedical
IIT-MTL-ClinicalBERT[e]	Wikipedia+BookCorpus+MIMIC-III[f]	Clinical

[a]MT-DNN: multi-task deep neural networks.

[b]RoBERTa: robustly optimized bidirectional encoder representations from transformers approach.

[c]BioBERT: bidirectional encoder representations from transformers for biomedical text mining.

[d]PMC: PubMed Central

[e]IIT-MTL-ClinicalBERT: iteratively trained using multi-task learning on ClinicalBERT.

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

Other Similarity Features

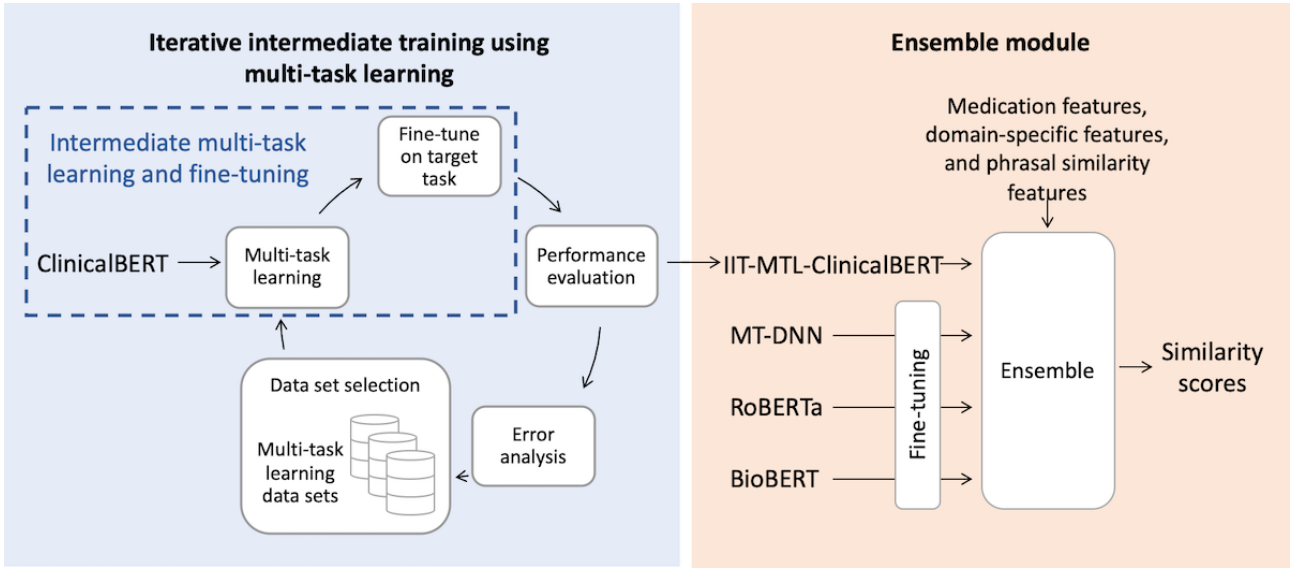
Under the hypothesis that aggregating similarity metrics from different perspectives could help further boost performance, we incorporated additional string similarity features to our ensembled model. On the basis of the observation that medication instructions appear frequently in our data set, we incorporated medication features by (1) using a medication information extraction system [69] to extract medications and its related attributes (eg, drug name, dosage, duration, form, frequency, route, and strength) from the text and (2) converting the extracted attributes into composite features. We also incorporated additional features shown to be useful in the previous 2018 ClinicalSTS challenge, including domain-specific features and phrasal similarity features. Details on these features are provided in Multimedia Appendix 3 [50,51,69-71].

Ensemble Methods

A total of 3 learning algorithms for regression were used for ensembling language model outputs and features: linear regression, Bayesian regression, and ridge regression. Note that we also explored random forest and XGBoost, which were used in the previous year's winning systems, but found that they underperformed, and therefore, we did not use those methods. On the basis of the performance on the internal test data set, we experimented with incrementally averaging different combinations of the constituent model outputs while adding the other similarity features previously described. A detailed explanation of the training parameters is provided in Multimedia Appendix 2.

Figure 4 presents an overview of our end-to-end system on the ClinicalSTS task, consisting of an iterative intermediate multi-task training step followed by an ensemble module. Note that the intermediate MTL and fine-tuning portion of Figure 4 was presented earlier in more detail in Figure 3.

Figure 4. Overview of our end-to-end system. ClinicalBERT: bidirectional encoder representations from transformers on clinical text; IIT-MTL-ClinicalBERT: iterative intermediate training using multi-task learning on ClinicalBERT; MT-DNN: multi-task deep neural networks; RoBERTa: robustly optimized BERT approach; BioBERT: bidirectional encoder representations from transformers for biomedical text mining.



Evaluation Metrics

We evaluated the proposed system using the evaluation script released by the organizers of the 2019 n2c2/OHNLP challenge to measure the Pearson correlation coefficient (PCC) between the human-annotated (gold standard) and predicted clinical semantic similarity scores. In the Results section, we report the PCC on the internal test data set for each iteration in IIT-MTL as well as on each combination of language models tried during ensembling. We also report the PCC for our 3 official submissions to the 2019 n2c2/OHNLP challenge on both the internal test data set and withheld external test data set.

Results

Iterative Intermediate Training Using MTL

Table 4 presents the results of each iteration in IIT-MTL. In comparison with the ClinicalBERT baseline, the addition of complementary data sets improved the overall model performance. Notably, not all data set additions resulted in improved performance. This is highlighted in iteration 5, where the addition of QQP led to a significant drop in performance. As the model from iteration 4 showed the best performance on the internal test data set, we adopted this variant for the final IIT-MTL-ClinicalBERT model.

Table 4. Results of each iteration of iterative intermediate training using multi-task learning.

Experiment and language model	Data sets used for iterative intermediate training approach using multi-task learning						Pearson correlation coefficient on internal test
	STS-B[a]	RQE[b]	MedNLI[c]	Topic	MedNER[d]	QQP[e]	
BL[f]							
1 BERT[g]	—[h]	—	—	—	—	—	0.834
2 ClinicalBERT[i]	—	—	—	—	—	—	0.848
Iter[j]							
1 ClinicalBERT	✓[k]	—	—	—	—	—	0.852
2 ClinicalBERT	✓	✓	✓	—	—	—	0.862
3 ClinicalBERT	✓	✓	✓	✓	—	—	0.866
4 ClinicalBERT	✓	✓	✓	✓	✓	—	*0.870*[l]
5 ClinicalBERT	✓	✓	✓	✓	✓	✓	0.856

[a]STS-B: semantic textual similarity benchmark.

[b]RQE: Recognizing Question Entailment.

[c]MedNLI: Natural Language Inference data set for the clinical domain.

[d]MedNER: Medication-NER data set.

[e]QQP: Quora Question Pair data set.

[f]BL: baseline.

[g]BERT: bidirectional encoder representations from transformers.

[h]Indicates data set was not used for this experiment.

[i]ClinicalBERT: bidirectional encoder representations from transformers on clinical text mining.

[j]Iter: iteration.

[k]Indicates data sets that were trained together in multi-task learning.

[l]Italics signify highest Pearson correlation coefficient obtained on internal test data set.

Ensemble Module

Table 5 presents the results of the language model ensemble experiments performed on the internal test data set. Here, the statistical mean of the normalized language model outputs was used as our ensemble method. Of the individual models, IIT-MTL-ClinicalBERT and BioBERT, which were pretrained on clinical and biomedical corpora, respectively, achieved higher PCC as compared with MT-DNN and RoBERTa, which were pretrained only on general domain corpora. In general, ensembled models performed better than the individual constituent models alone, with the combination of IIT-MTL-ClinicalBERT, BioBERT, and MT-DNN resulting in the highest performance (PCC 0.8809) on the internal test data set.

Table 5. Ablation study of language models utilized in the ensemble module. The statistical mean of the language model outputs was used as the ensembling method.

Experiment	Language model ensemble				Pearson correlation coefficient on internal test
	IIT-MTL-ClinicalBERT[a]	BioBERT[b]	MT-DNN[c]	RoBERTa[d]	
1	✓[e]	—[f]	—	—	0.8711
2	—	✓	—	—	0.8707
3	—	—	✓	—	0.8685
4	—	—	—	✓	0.8578
5	✓	✓	—	—	0.8754
6	—	✓	✓	—	0.8780
7	—	—	✓	✓	0.8722
8	✓	—	—	✓	0.8741
9	✓	—	✓	—	0.8796
10	—	✓	—	✓	0.8720
11	✓	✓	✓	—	*0.8809* [g]
12	—	✓	✓	✓	0.8769
13	✓	—	✓	✓	0.8787
14	✓	✓	—	✓	0.8764
15	✓	✓	✓	✓	0.8795

[a]IIT-MTL-ClinicalBERT: iterative intermediate training using multi-task learning on ClinicalBERT.

[b]BioBERT: bidirectional encoder representations from transformers for biomedical text mining.

[c]MT-DNN: multi-task deep neural networks.

[d]RoBERTa: robustly optimized bidirectional encoder representations from transformers approach.

[e]Indicates which language models are included in the ensemble.

[f]Indicates language model was not used for this experiment.

[g]Italics signify the highest Pearson correlation coefficient obtained on internal test data set.

On the basis of the experiments presented in Table 5, IIT-MTL-ClinicalBERT & BioBERT & MT-DNN was adopted as the base combination of language models for our official submissions. Table 6 presents the results of this base combination of language models, with incremental addition of other similarity features using four different ensemble methods. Results are shown for both the internal and withheld external test data sets. Note that the addition of domain-specific and phrasal similarity features has been included in Table 6 for completeness (although it resulted in lower performance) because it was part of our official submissions.

Table 6. End-to-end ensemble module and official submission results.

Components	Pearson correlation coefficient on internal test[a]				Pearson correlation coefficient on external test[a]			
	Mean	LR[b]	BR[c]	RR[d]	Mean	LR	BR	RR
IIT-MTL-ClinicalBERT[e] & MT-DNN[f] & BioBERT[g]	*0.8809*	0.8796	0.8795	0.8796	*0.9006*	0.8978	0.8978	0.8978
+ medication features	N/A[h]	*0.8841*	0.8832	0.8831	N/A	*0.9010*	0.8997	0.8975
+ domain-specific and phrasal similarity features	N/A	0.8733	0.8741	*0.8799*	N/A	0.8861	0.8920	*0.8875*

[a]Italics signify the Pearson correlation coefficient obtained on the internal and external test data set corresponding to the three configurations (components and ensemble method) that were our official submissions to the 2019 n2c2/OHNLP challenge.

[b]LR: linear regression.

[c]BR: Bayesian regression.

[d]RR: ridge regression.

[e]IIT-MTL-ClinicalBERT: iterative intermediate training using multi-task learning on ClinicalBERT.

[f]MT-DNN: multi-task deep neural networks.

[g]BioBERT: bidirectional encoder representations from transformers for biomedical text mining.

[h]N/A: not applicable.

Official Submission

The best performing configurations on the internal test data set, as shown in Table 6, were entered as our official submissions to the 2019 n2c2/OHNLP ClinicalSTS challenge. The details of each of our 3 official submissions are as follows:

- Submission 1: IIT-MTL-ClinicalBERT & MT-DNN & BioBERT
  - A statistical mean of the scores produced by the language models, specifically IIT-MTL-ClinicalBERT, MT-DNN, and BioBERT.
- Submission 2: IIT-MTL-ClinicalBERT & MT-DNN & BioBERT+medication features
  - A linear regression model trained on each component output from Submission 1 and medication features.
- Submission 3: IIT-MTL-ClinicalBERT & MT-DNN & BioBERT+medication features+domain-specific and phrasal similarity features
  - A ridge regression model trained on all features from Submission 2 and phrasal similarity and domain-specific features.

Our submission 2 achieved first place out of 87 submitted systems with a PCC of 0.9010 based on the official results. Our submission 1 achieved second place with a PCC of 0.9006.

With the release of the external test data set, we reran the experiments for language model ensembling on the external test data set. We identified the highest performing configuration on the external test data set as the statistical mean of the scores produced by the combination of IIT-MTL-ClinicalBERT, MT-DNN, and RoBERTa, which resulted in a PCC of 0.9025.

Discussion

Principal Findings

Iterative intermediate training using MTL is an effective way to leverage annotated data from related tasks to improve performance on the target task. However, it is critical to select data sets that can induce contextualized embeddings necessary for the target task. If the network is tasked with making predictions on unrelated tasks, negative transfer may ensue, resulting in lower quality predictions on the target task. Applying IIT-MTL to train ClinicalBERT with related tasks—STS-B, RQE, MedNLI, Topic, and MedNER—resulted in improved performance on the target ClinicalSTS task. However, the addition of QQP to the MTL step resulted in a significant drop in performance. This may be attributed to the fact that, in contrast to the other data sets used, QQP was created for a different sentence pair task (classification rather than regression) on the general domain (as opposed to RQE and MedNLI, which are on the clinical domain). This illustrates the importance of data set selection for the effectiveness of the intermediate multi-task training step.

Ensembling language models pretrained on domain-specific and domain-independent corpora incorporates different aspects of clinical semantic similarity. Table 7 presents the ground truth for two sentence pairs, along with predictions from each constituent model. The first sentence pair contains minimal domain-specific terminology; hence, the models trained on domain-independent corpora, MT-DNN and RoBERTa, predicted scores closer to the ground truth. The low ground truth score in the second sentence pair is because of dissimilar clinical concepts within the text; hence, the models trained on domain-specific corpora, IIT-MTL-ClinicalBERT and BioBERT, predicted scores closer to the ground truth.

Table 7. Sample sentence pairs with ground truth annotations and predictions from three language models used in the final ensembled system.

Sentence 1	Sentence 2	Ground Truth	Predictions			
			IIT-MTL-Clinical-BERT[a]	BioBERT[b]	MT-DNN[c]	RoBERTa[d]
"The following consent was read to the patient and accepted to order testing."	"We explained the risks, benefits, and alternatives, and the patient agreed to proceed."	2.5	0.61	1.01	2.15	2.51
"Negative for coughing up blood, coughing up mucus (phlegm) and wheezing."	"Negative for abdominal pain, blood in stool, constipation, diarrhea and vomiting."	0.5	1.04	1.18	2.34	1.74

[a]IIT-MTL-ClinicalBERT: iterative intermediate training using multi-task learning on ClinicalBERT.

[b]BioBERT: bidirectional encoder representations from transformers for biomedical text mining.

[c]MT-DNN: multi-task deep neural networks.

[d]RoBERTa: robustly optimized bidirectional encoder representations from transformers approach.

Analysis of Model Performance

Our best official submission achieved a PCC of 0.9010 on the external test data set. However, the model performance varies significantly depending on the gold similarity scores. On the low and high ends of the gold scores, [0-2) or [4-5], our model achieves a PCC of 0.9234. However, in the middle range of the gold scores, [2-4), it performs much worse with a PCC of 0.5631. The lower performance in the middle range can be partially attributed to ground truth issues. Weak-to-moderate interannotator agreement (0.6 weighted Cohen kappa) coupled with the lack of an adjudication process (scores from 2 annotators were averaged to provide the gold score), led to concentration of annotation errors in the middle range of the gold scores. For example, greater disagreement between 2 annotators (eg, gold scores 1 and 5) will end up in the middle range (final averaged score 3) as compared with low disagreements (eg, 4 and 5 with the final score of 4.5). The drop in performance in the middle range may also indicate that although our model performs well at distinguishing completely similar or dissimilar sentence pairs, it struggles in scoring sentences with moderate clinical semantic similarity.

To further investigate this behavior, we studied how predictions varied for each similarity interval using the withheld external test data set. For this, we converted the continuous range gold scores and our model predictions into 5 intervals: [0,1), [1-2), [2-3), [3-4), [4-5]. Using these intervals, we then calculated the F1-score by computing true positives, false positives, and false negatives. A prediction is a true positive if the gold score is in the same similarity interval as the prediction; otherwise, it is termed as false positive (in the predicted interval) and false negative (in the gold interval). Our best model achieves a relatively high F1-score at the extreme ranges (0.77, 0.80, and 0.71 for [0,1), [1-2), [4-5], respectively) but struggles in the middle intervals (0.23 and 0.44 for [2-3) and [3-4), respectively).

Limitations and Future Work

We acknowledge certain limitations of this study. First, these results are specific to the 2019 n2c2/OHNLP ClinicalSTS data set, which contains clinical text snippets from a single EHR data warehouse (Mayo Clinic EHR data warehouse). Furthermore, the chosen sentence pairs have high surface lexical similarity (ie, candidate pairs must have ≥0.45 average score of Ratcliff/Obershelp pattern matching algorithm, cosine similarity, and Levenshtein distance), which limits the variation in the data set. Thus, there is a need to validate this process on a more diverse ground truth, which (1) contains clinical text from multiple data warehouses and (2) allows for a less restrictive sentence pairing. Second, we observed inconsistencies in the ground truth, which may be inherent to a complex task such as clinical semantic textual similarity. We have made preliminary progress in quantifying these errors and their impact on the results, but more work is needed in this direction. Finally, although our system has achieved high PCC on the ClinicalSTS task, additional research is still needed to understand how to apply this foundational task to the real-world problem of bloated, disorganized clinical documentation.

Although our system achieved state-of-the-art results in the challenge, the proposed system has following avenues for improvement and further exploration:

1. The data set selection process in IIT-MTL is largely manual, driven by empirical observations and domain knowledge. Recent developments in automatic machine learning (AutoML), ranging from optimizing hyper-parameters using random search [72] to discovering novel neural architectures using reinforcement learning [73], have shown promising results. We plan to explore AutoML to relieve this manual effort in the future.
2. The language model ensemble works well for inducing domain-specific and domain-independent knowledge. However, this process remains largely intuitive. We plan to explore how language modeling objectives influence the domain adaptability of the learned language models on the target task.
3. At the time of the challenge, we applied our IIT-MTL methodology only to ClinicalBERT because of time constraints. We plan to employ our IIT-MTL methodology on other implemented language models and evaluate their performance.
4. Our proposed system has a significant computational cost, as we leverage several transformer-based language models. We plan to explore the performance impact of replacing

these models with their less computationally expensive counterparts [74].

5. In our experiments, inclusion of domain-specific and phrasal features led to a drop in performance. This is likely because of effective learning of these features by pretrained transformer-based language models, as observed in the general domain [75,76]. We wish to investigate this behavior further by utilizing probing tasks [77] in transformer language models.

Conclusions

In this study, we presented an effective methodology leveraging (1) an iterative intermediate training step in a MTL setup and (2) multiple language models pretrained on diverse corpora, which achieved first place in the 2019 ClinicalSTS challenge. This study demonstrates the potential for IIT-MTL to improve the performance of other tasks restricted by limited data sets. This contribution opens new avenues of exploration for optimized data set selection to generate more robust and universal contextual representations of text in the clinical domain.

Acknowledgments

The authors wish to thank Dr Bharath Dandala and Venkata Joopudi for providing valuable feedback on the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data sets used in iterative intermediate training approach using multi-task learning methodology.
[PDF File (Adobe PDF File), 159 KB - medinform_v8i11e22508_app1.pdf]

Multimedia Appendix 2
Experimental settings.
[PDF File (Adobe PDF File), 159 KB - medinform_v8i11e22508_app2.pdf]

Multimedia Appendix 3
Implementation details of other similarity features.
[PDF File (Adobe PDF File), 86 KB - medinform_v8i11e22508_app3.pdf]

References

1. Jamoom EW, Patel V, Furukawa MF, King J. EHR adopters vs non-adopters: impacts of, barriers to, and federal initiatives for EHR adoption. Healthc (Amst) 2014 Mar;2(1):33-39 [FREE Full text] [doi: 10.1016/j.hjdsi.2013.12.004] [Medline: 26250087]
2. Zhang R, Pakhomov S, McInnes BT, Melton GB. Evaluating measures of redundancy in clinical texts. AMIA Annu Symp Proc 2011;2011:1612-1620 [FREE Full text] [Medline: 22195227]
3. Shoolin J, Ozeran L, Hamann C, Bria W. Association of medical directors of information systems consensus on inpatient electronic health record documentation. Appl Clin Inform 2013;4(2):293-303 [FREE Full text] [doi: 10.4338/ACI-2013-02-R-0012] [Medline: 23874365]
4. Vogel L. Cut-and-paste clinical notes confuse care, say US internists. Can Med Assoc J 2013 Dec 10;185(18):E826 [FREE Full text] [doi: 10.1503/cmaj.109-4656] [Medline: 24218539]
5. Dimick C. Documentation bad habits. Shortcuts in electronic records pose risk. J AHIMA 2008 Jun;79(6):40-43. [Medline: 18604974]
6. Wang MD, Khanna R, Najafi N. Characterizing the source of text in electronic health record progress notes. JAMA Intern Med 2017 Aug 1;177(8):1212-1213 [FREE Full text] [doi: 10.1001/jamainternmed.2017.1548] [Medline: 28558106]
7. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 2;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
8. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010 Mar 1;2010:1-5 [FREE Full text] [Medline: 21347133]
9. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resour Eval 2018 Oct 24;54(1):57-72. [doi: 10.1007/s10579-018-9431-1]
10. Yanshan W, Sunyang F, Feichen S, Sam H, Ozlem UH. Overview of the 2019 N2C2/OHNLP track on clinical semantic textual similarity. JMIR Med Informatics Preprint posted online August 10, 2020. [FREE Full text] [doi: 10.2196/preprints.23375]

11. Wang Y, Rastegar-Mojarad M, Afzal N, Liu S, Wang L, Shen F, et al. Overview of BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. Clin Semantic Text Sim Preprint posted online August 2018. [FREE Full text] [doi: 10.13140/RG.2.2.26682.24006]
12. Rastegar-Mojarad M, Liu S, Wang Y, Afzal N, Wang L, Shen F, et al. BioCreative/OHNLP Challenge 2018. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2018 Aug Presented at: ACM-BCB'18; August 29-September 1, 2018; Washington, DC. [doi: 10.1145/3233547.3233672]
13. Agirre E, Cer D, Diab M, Gonzalez-Agirre A. Task 6: A Pilot on Semantic Textual Similarity. In: First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (Semeval 2012). 2012 Presented at: SEM'12; June 7-8, 2012; Montreal, Canada p. 385-393. [doi: 10.18653/v1/s15-2045]
14. Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. Shared task: Semantic Textual Similarity. In: Second Joint Conference on Lexical and Computational Semantics. 2013 Presented at: SEM'13; June 13-14, 2013; Atlanta, Georgia, USA p. 32-43.
15. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. Task 10: Multilingual Semantic Textual Similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation. 2014 Presented at: SemEval'14; August 23-24, 2014; Dublin, Ireland p. 81-91. [doi: 10.3115/v1/s14-2010]
16. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation. 2015 Presented at: SemEval'15; June 4-5, 2015; Denver, Colorado p. 252-263. [doi: 10.18653/v1/s15-2045]
17. Agirre E, Banea C, Cer D, Diab M, Gonzalez-Agirre A, Mihalcea R, et al. Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation. 2016 Presented at: SemEval'16; June 16-17, 2016; San Diego, California p. 497-511. [doi: 10.18653/v1/s16-1081]
18. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation. 2017 Presented at: SemEval'17; August 3-4, 2017; Vancouver, Canada p. 1-14. [doi: 10.18653/v1/s17-2001]
19. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Presented at: NAACL HLT'19; June 2-7, 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/N19-1423]
20. Phang J, Févry T, Bowman S. Sentence encoders on STILTs: supplementary training on intermediate labeled-data tasks. arXiv 2018 epub ahead of print [FREE Full text]
21. Liu X, He P, Chen W, Gao J. Multi-Task Deep Neural Networks for Natural Language Understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL'19; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-1441]
22. Dolan W, Brockett C. Automatically Constructing a Corpus of Sentential Paraphrases. In: Third International Workshop on Paraphrasing. 2005 Presented at: IWP'05; October 11-13, 2005; Jeju Island, Korea URL: https://www.aclweb.org/anthology/I05-5002.pdf
23. Williams A, Nangia N, Bowman S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018 Presented at: NAACL-HLT'18; June 1-6, 2018; New Orleans, Louisiana. [doi: 10.18653/v1/n18-1101]
24. Romanov A, Shivade C. Lessons from Natural Language Inference in the Clinical Domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP'18; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-1187]
25. Bowman S, Angeli G, Potts C, Manning C. A Large Annotated Corpus for Learning Natural Language Inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Presented at: EMNLP'15; September 17-21, 2015; Lisbon, Portugal. [doi: 10.18653/v1/d15-1075]
26. Saríc F, Glavaš G, Karan M, Šnajder J, Bašić B. TakeLab: Systems for Measuring Semantic Text Similarity. In: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. 2012 Presented at: SEM'12; June 7-8, 2012; Montreal, Canada.
27. Jimenez S, Becerra C, Gelbukh A. Soft Cardinality: A Parameterized Similarity Function for Text Comparison. In: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. 2012 Presented at: SEM'12; June 7-8, 2012; Montreal, Canada.
28. Bär D, Biemann C, Gurevych I, Zesch T. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. 2012 Presented at: SEM'12; June 7-8, 2012; Montreal, Canada.

29. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. 2013 Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, Nevada p. 3111-3119. [doi: 10.5555/2999792.2999959]
30. Hanson E. Musicassette interchangeability: the facts behind the facts. AES J Audio Eng Soc 1971;19(5):- [FREE Full text]
31. Wieting J, Bansal M, Gimpel K, Livescu K. From paraphrase database to compositional paraphrase model and back. Trans Assoc Comput Linguist 2015 Dec;3:345-358. [doi: 10.1162/tacl_a_00143]
32. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017 Presented at: EACL'17; April 3-7, 2017; Valencia, Spain. [doi: 10.18653/v1/e17-2068]
33. Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: Proceedings of the 31st International Conference on Machine Learning. 2014 Presented at: ICML'14; June 21–26, 2014; Beijing, China.
34. Lau J, Baldwin T. An Empirical Evaluation of DOC2VEC with Practical Insights into Document Embedding Generation. In: Proceedings of the 1st Workshop on Representation Learning for NLP. 2016 Aug Presented at: REPL4NLP'16; August 11, 2016; Berlin, Germany p. 78-86. [doi: 10.18653/v1/w16-1609]
35. Pagliardini M, Gupta P, Jaggi M. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018 Presented at: NAACL-HLT'18; June 1-6, 2018; New Orleans, Louisiana. [doi: 10.18653/v1/n18-1049]
36. Conneau A, Kiela D, Schwenk H, Barrault L. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: EMNLP'17; September 7–11, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/d17-1070]
37. Arora S, Liang Y, Ma T. A Simple but Tough-to-beat Baseline for Sentence Embeddings. In: 5th International Conference on Learning Representations. 2017 Presented at: ICLR'17; April 24-26, 2017; Toulon, France.
38. Shao Y. Task 1: Use convolutional neural network to evaluate Semantic Textual Similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation. 2017 Presented at: SemEval'17; August 3-4, 2017; Vancouver, Canada. [doi: 10.18653/v1/s17-2016]
39. Huang PS, He X, Gao J, Deng L, Acero A, Heck L. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. 2013 Presented at: CIKM'13; October 1, 2016; San Franciso, California. [doi: 10.1145/2505515.2505665]
40. Howard J, Ruder S. Universal Language Model Fine-Tuning for Text Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 Presented at: ACL'18; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1031]
41. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. Semantic Scholar. 2018. URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2020-11-02]
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA p. 2017. [doi: 10.5555/3295222.3295349]
43. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: International Conference on Learning Representations. 2019 Presented at: ICLR'19; May 6-9, 2019; New Orleans.
44. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
45. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arxiv 2019:- epub ahead of print [FREE Full text]
46. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP'19; June 7, 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/w19-1909]
47. Zhang Y, Yang Q. A survey on multi-task learning. arXiv 2017:- epub ahead of print [FREE Full text]
48. Chen Q, Du J, Kim S, Wilbur WJ, Lu Z. Combining Rich Features and Deep Learning for Finding Similar Sentences in Electronic Medical Records. In: Proceedings of the BioCreative/OHNLP Challenge. 2018 Presented at: OHNLP'18; September 1-8, 2018; Washington, DC URL: https://www.researchgate.net/publication/327402060_Combining_rich_features_and_deep_learning_for_finding_similar_sentences_in_electronic_medical_records
49. Tian J, Zhou Z, Lan M, Wu Y. Task 1: Leverage Kernel-based Traditional NLP features and Neural Networks to Build a Universal Model for Multilingual and Cross-lingual Semantic Textual Similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation. 2017 Presented at: SemEval'17; August 3-4, 2017; Vancouver, Canada. [doi: 10.18653/v1/s17-2028]

50. Sultan M, Bethard S, Sumner T. DLSCU: Sentence Similarity from Word Alignment. In: Proceedings of the 8th International Workshop on Semantic Evaluation. 2014 Presented at: SemEval'14; August 23-24, 2014; Denver, Colorado. [doi: 10.3115/v1/s14-2039]
51. Sultan M, Bethard S, Sumner T. DLSCU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In: Proceedings of the 9th International Workshop on Semantic Evaluation. 2015 Presented at: SemEval'15; June 4-5, 2015; Denver, Colorado. [doi: 10.18653/v1/s15-2027]
52. Cer D, Yang Y, Kong S, Hua N, Limtiaco N, John RS. Universal sentence encoder. arXiv 2018:- epub ahead of print [FREE Full text] [doi: 10.18653/v1/d18-2029]
53. Mulyar A, McInnes B. MT-clinical BERT: scaling clinical information extraction with multitask learning. arXiv 2020:- [FREE Full text]
54. Peng Y, Chen Q, Lu Z. An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. 2020 Presented at: BioNLP'20; July 9, 2020; Online. [doi: 10.18653/v1/2020.bionlp-1.22]
55. Ratcliff/Obershelp Pattern Recognition. NIST: National Institute of Standards and Technology. 2004. URL: https://xlinux.nist.gov/dads/HTML/ratcliffObershelp.html [accessed 2020-11-01]
56. Li D, Rastegar-Mojarad M, Elayavilli R, Wang Y, Mehrabi S, Yu Y, et al. A Frequency-filtering Strategy of Obtaining PHI-free Sentences From Clinical Data Repository. In: Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. 2015 Presented at: BCB'15; September 9–12, 2015; Atlanta, GA. [doi: 10.1145/2808719.2808752]
57. Abacha AB, Dina D. Recognizing Question Entailment for Medical Question Answering. In: Proceedings of the Annual Symposium. 2016 Presented at: AMIA'16; June 12-18, 2016; Chicago, Illinois.
58. Sharma L, Graesser L, Nangia N, Evci U. Natural language understanding with the quora question pairs dataset. arXiv 2019 epub ahead of print [FREE Full text]
59. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
60. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010;17(1):19-24 [FREE Full text] [doi: 10.1197/jamia.M3378] [Medline: 20064797]
61. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. Br Med J 2000 Aug 12;321(7258):429-432 [FREE Full text] [doi: 10.1136/bmj.321.7258.429] [Medline: 10938054]
62. Quora Question Pairs. Kaggle. URL: https://www.kaggle.com/c/quora-question-pairs [accessed 2020-11-02]
63. Wu Y, Schuster M, Chen Z, Le QV. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv 2016 epub ahead of print [FREE Full text]
64. namisan/mt-dnn: Multi-Task Deep Neural Networks for Natural Language Understanding. GitHub. URL: https://github.com/namisan/mt-dnn [accessed 2020-11-02]
65. International Conference on Learning Representations ICLR. 2015. URL: https://arxiv.org/pdf/1412.6980.pdf [accessed 2020-11-02]
66. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized bert pretraining approach. arXiv. 2019. URL: http://arxiv.org/abs/1907.11692 [accessed 2020-11-01]
67. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016 Presented at: EMNLP'16; November 1-5, 2016; Austin, Texas. [doi: 10.18653/v1/d16-1264]
68. Dagan I, Glickman O, Magnini B. The PASCAL Recognising Textual Entailment Challenge. Berlin, Heidelberg: Springer; 2006.
69. Mahajan D, Liang J, Tsou C. Extracting Daily Dosage from Medication Instructions in EHRs: An Automated Approach and Lessons Learned. arXiv org. 2020. URL: https://arxiv.org/pdf/2005.10899.pdf [accessed 2020-11-01]
70. Lindberg D, Humphreys B, McCray A. The unified medical language system. Methods Inf Med 2018 Feb 06;32(04):281-291. [doi: 10.1055/s-0038-1634945]
71. Miller GA. WordNet: a lexical database for English. Commun ACM 1995 Nov;38(11):39-41. [doi: 10.1145/219717.219748]
72. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-parameter Optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011 Presented at: NIPS'11; December 12-14, 2011; Granada, Spain. [doi: 10.5555/2986459.2986743]
73. Zhong Z, Yan J, Wu W, Shao J, Liu C. Practical Block-Wise Neural Network Architecture Generation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018 Presented at: IEEE'18; June 18-22, 2018; Salt Lake City, UT. [doi: 10.1109/cvpr.2018.00257]
74. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT. arXiv 2019;2 epub ahead of print [FREE Full text]

75. Tenney I, Das D, Pavlick E. BERT Rediscovers the Classical NLP Pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL'19; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-1452]
76. Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology. arXiv 2020 epub ahead of print [FREE Full text]
77. Tenney I, Xia P, Chen B, Wang A, Poliak A, Thomas MR, et al. What Do You Learn From Context? Probing for Sentence Structure in Contextualized Word Representations. In: Seventh International Conference on Learning Representations. 2019 Presented at: ICLR'19; May 6-9, 2019; New Orleans.

Abbreviations

BERT: bidirectional encoder representations from transformers
BioBERT: bidirectional encoder representations from transformers for biomedical text mining
ClinicalBERT: bidirectional encoder representations from transformers on clinical text mining
ClinicalSTS: clinical semantic textual similarity
EHR: electronic health record
IIT-MTL-ClinicalBERT: iterative intermediate training using multi-task learning on ClinicalBERT
IIT-MTL: iterative intermediate training approach using multi-task learning
MedNER: medication named entity recognition data set
MedNLI: natural language inference data set for the clinical domain
MIMIC-III: Medical Information Mart for Intensive Care III
MT-DNN: multi-task deep neural networks
MTL: multi-task learning
n2c2: National Natural Language Processing Clinical Challenges
NLP: natural language processing
OHNLP: Open Health Natural Language Processing Consortium
PCC: Pearson correlation coefficient
PMC: PubMed Central
QQP: Quora question pairs
RoBERTa: robustly optimized bidirectional encoder representations from transformers approach
RQE: recognizing question entailment
STS-B: semantic textual similarity benchmark
STS: semantic textual similarity

Edited by Y Wang; submitted 31.07.20; peer-reviewed by K Verspoor, R Abeysinghe, TL Sun; comments to author 22.09.20; revised version received 10.10.20; accepted 13.10.20; published 27.11.20.

Please cite as:
Mahajan D, Poddar A, Liang JJ, Lin YT, Prager JM, Suryanarayanan P, Raghavan P, Tsou CH
Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning
JMIR Med Inform 2020;8(11):e22508
URL: http://medinform.jmir.org/2020/11/e22508/
doi:10.2196/22508
PMID:33245284

Original Paper

The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview

Yanshan Wang[1], PhD; Sunyang Fu[1], MHI; Feichen Shen[1], PhD; Sam Henry[2], PhD; Ozlem Uzuner[2], PhD; Hongfang Liu[1], PhD

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

[2]Information Sciences and Technology, George Mason University, Fairfax, VA, United States

Corresponding Author:
Yanshan Wang, PhD
Department of Health Sciences Research
Mayo Clinic
200 1st Street SW
Rochester, MN, 55905
United States
Phone: 1 507293138
Email: wang.yanshan@mayo.edu

Abstract

Background: Semantic textual similarity is a common task in the general English domain to assess the degree to which the underlying semantics of 2 text segments are equivalent to each other. Clinical Semantic Textual Similarity (ClinicalSTS) is the semantic textual similarity task in the clinical domain that attempts to measure the degree of semantic equivalence between 2 snippets of clinical text. Due to the frequent use of templates in the Electronic Health Record system, a large amount of redundant text exists in clinical notes, making ClinicalSTS crucial for the secondary use of clinical text in downstream clinical natural language processing applications, such as clinical text summarization, clinical semantics extraction, and clinical information retrieval.

Objective: Our objective was to release ClinicalSTS data sets and to motivate natural language processing and biomedical informatics communities to tackle semantic text similarity tasks in the clinical domain.

Methods: We organized the first BioCreative/OHNLP ClinicalSTS shared task in 2018 by making available a real-world ClinicalSTS data set. We continued the shared task in 2019 in collaboration with National NLP Clinical Challenges (n2c2) and the Open Health Natural Language Processing (OHNLP) consortium and organized the 2019 n2c2/OHNLP ClinicalSTS track. We released a larger ClinicalSTS data set comprising 1642 clinical sentence pairs, including 1068 pairs from the 2018 shared task and 1006 new pairs from 2 electronic health record systems, GE and Epic. We released 80% (1642/2054) of the data to participating teams to develop and fine-tune the semantic textual similarity systems and used the remaining 20% (412/2054) as blind testing to evaluate their systems. The workshop was held in conjunction with the American Medical Informatics Association 2019 Annual Symposium.

Results: Of the 78 international teams that signed on to the n2c2/OHNLP ClinicalSTS shared task, 33 produced a total of 87 valid system submissions. The top 3 systems were generated by IBM Research, the National Center for Biotechnology Information, and the University of Florida, with Pearson correlations of r=.9010, r=.8967, and r=.8864, respectively. Most top-performing systems used state-of-the-art neural language models, such as BERT and XLNet, and state-of-the-art training schemas in deep learning, such as pretraining and fine-tuning schema, and multitask learning. Overall, the participating systems performed better on the Epic sentence pairs than on the GE sentence pairs, despite a much larger portion of the training data being GE sentence pairs.

Conclusions: The 2019 n2c2/OHNLP ClinicalSTS shared task focused on computing semantic similarity for clinical text sentences generated from clinical notes in the real world. It attracted a large number of international teams. The ClinicalSTS shared task could continue to serve as a venue for researchers in natural language processing and medical informatics communities to develop and improve semantic textual similarity techniques for clinical text.

(JMIR Med Inform 2020;8(11):e23375) doi:10.2196/23375

KEYWORDS

natural language processing; clinical natural language processing; medical natural language processing; semantic textual similarity; ClinicalSTS; n2c2; electronic health records; challenge; shared task

Introduction

Background

Semantic textual similarity (STS) is a common task in the general English domain to assess the degree to which the underlying semantics of 2 segments of text are equivalent to each other. Equivalency is usually assessed using ordinal scaled output ranging from complete semantic equivalence to complete semantic dissimilarity. Applications of STS include machine translation, summarization, text generation, question answering, short answer grading, semantic search, and dialogue and conversational systems.

Clinical Semantic Textual Similarity (ClinicalSTS) is the application of STS techniques in the clinical domain that attempts to measure the degree of semantic equivalence between 2 snippets of clinical text. Due to the wide adoption of electronic health record (EHR) systems, a vast volume of free-text EHR data has been generated [1], such as progress notes, discharge summaries, radiology reports, and pathology reports. The frequent use of copy and paste, templates, and smart phrases (eg, one can type a few characters that automatically expand to a longer phrase or template) has resulted in redundancy in clinical text. This reduces the quality of EHR data and adds to the cognitive burden of tracking complex medical records in clinical practice [2]. An analysis of 23,630 progress notes written by 460 clinicians showed that 18% of the text was manually entered, 46% was copied, and 36% was imported [3].

Studies that evaluated and measured redundancy in clinical text [2] showed that STS techniques are rarely applied in the clinical domain to reduce redundancy. ClinicalSTS can identify redundant clinical sentences, that is, semantically equivalent clinical texts, by computing the similarity score between 2 clinical snippets. Removing those redundant clinical sentences is vital to many clinical applications, such as clinical text summarization, clinical semantic information retrieval, and clinical decision support systems [4].

The STS shared task has been held annually since 2012 to encourage and support research in this area [5-10]. However, STS techniques have been rarely studied on clinical texts, and to our knowledge there are no clinical STS shared tasks. To motivate natural language processing (NLP) and biomedical informatics communities to study STS problems in the clinical domain, we organized the first ClinicalSTS challenge, the BioCreative/OHNLP ClinicalSTS shared task, in 2018 [11] to provide a venue for the evaluation of state-of-the-art algorithms and models by making available a real-world clinical note data set. The shared task attracted 4 participating teams that produced a total of 12 system submissions [12].

Objective

In 2019, we continued the shared task as a collaboration with National NLP Clinical Challenges (n2c2) and the Open Health Natural Language Processing (OHNLP) consortium under the name n2c2/OHNLP track on ClinicalSTS [11]. Our aim was for the community to tackle STS problems in the clinical domain in a workshop at the American Medical Informatics Association 2019 Annual Symposium. In this paper, we first give an overview of the ClinicalSTS task and how we prepared the data set for the 2019 shared task differently from that in the previous year. Then, we describe the record number of participating teams and their systems. Finally, we present the results, system rankings, and future research directions for the ClinicalSTS task.

Methods

Task Overview

ClinicalSTS provides paired clinical text snippets for each participant. The clinical text snippets are mostly sentences extracted from clinical notes. The participating systems are asked to return a numerical score indicating the degree of semantic similarity between the 2 sentences. Performance is measured by the Pearson correlation coefficient between the predicted similarity scores and human judgments. The ClinicalSTS scores fall on an ordinal scale, ranging from 0 to 5, where 0 means that the 2 clinical text snippets are completely dissimilar (ie, no overlap in their meanings) and 5 means that the 2 snippets have complete semantic equivalence. Our previous publications [12,13] showed clinical text examples of the ordinal similarity scale. Participating systems can use real valued scores to indicate their semantic similarity prediction.

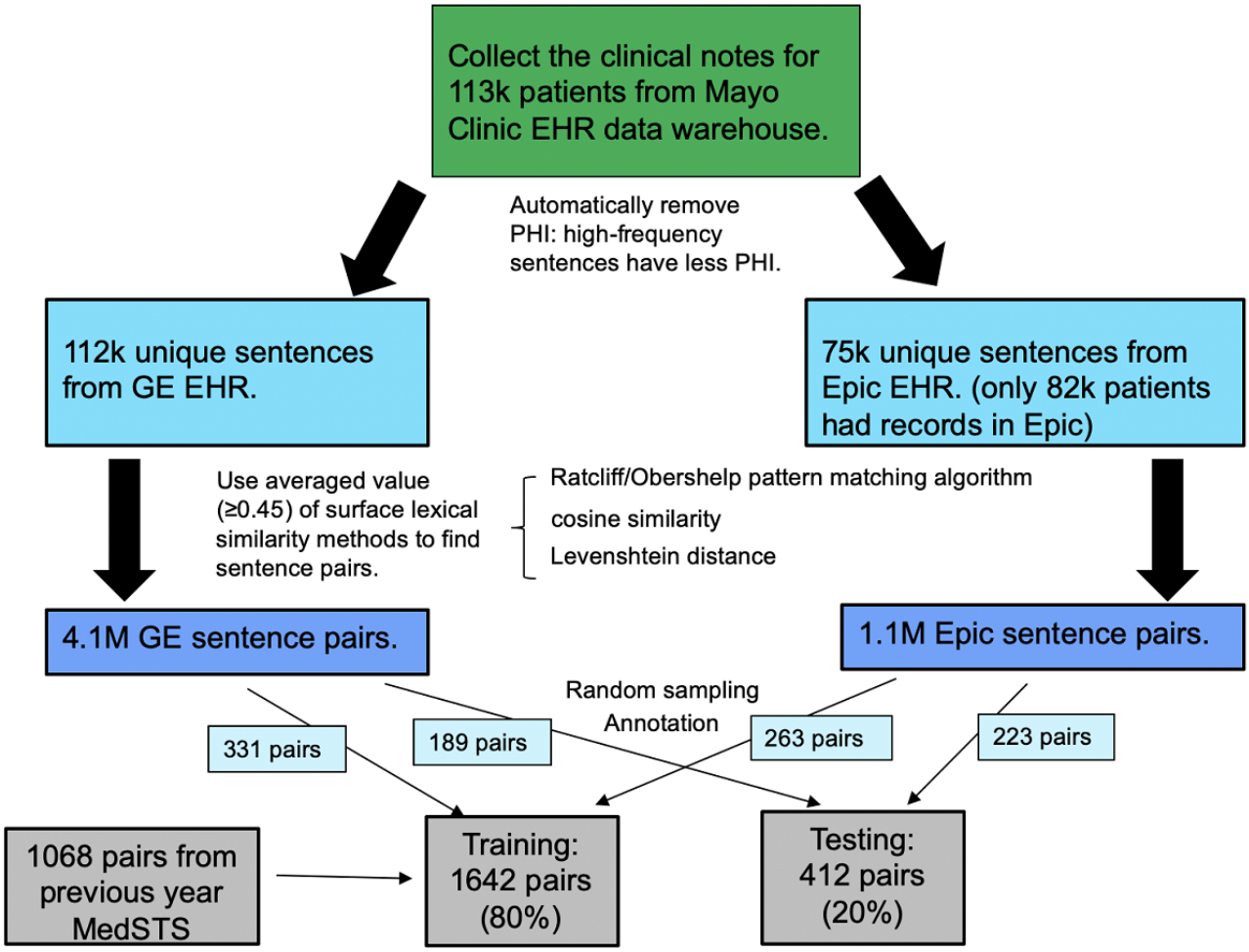
Data Preparation

We collected the data set for the 2019 ClinicalSTS shared task from EHRs at the Mayo Clinic's clinical data warehouse. Both the study and a waiver of informed consent were approved by the Mayo Clinic Institutional Review Board in accordance with 45 CFR 46.116 (approval no. 17–003030). Since the Mayo Clinic had completed a systemwide EHR transition across all care sites from GE Healthcare to Epic Systems Corporation, the data set in the 2019 shared task combined the data set from the 2018 shared task, which was an annotated subset of the MedSTS data set [13], and a new data set extracted from the historical GE EHR system and Epic EHR system. By combining data sets, we aimed to compare the semantics in clinical text generated from 2 different EHR systems. We did not release the EHR source information to the participating teams during the shared task.

Figure 1 illustrates the data set used for this shared task. To curate the data set, we first collected clinical notes from the clinical data warehouse for 113,000 patients receiving their primary care at the Mayo Clinic. We removed protected health information (PHI) by employing a frequency filtering approach [14] based on the assumption that sentences appearing in multiple patients' records tend to contain no PHI, which resulted in 112,00 unique sentences from the GE and 75,000 unique sentences from the Epic EHRs. We used the averaged value

(≥0.45) of 3 surface lexical similarities, namely the Ratcliff/Obershelp pattern-matching algorithm [15], cosine similarity [16], and Levenshtein distance [17], as a cutoff value to obtain candidate sentence pairs with some level of prima facie similarity. Wang et al [13] details how these methods were employed. We obtained 4.1 million GE sentence pairs and 1.1 million Epic sentence pairs. We randomly selected 1006 sentence pairs to be annotated by human experts. To ensure that no PHI existed in the final released data set, we manually removed PHI from each sentence. In the annotation phase, we asked 2 clinical experts to independently annotate each sentence pair in the ClinicalSTS data set on the basis of their semantic equivalence. Both annotators were very knowledgeable and had many years of experience in the clinical domain. Agreement between the 2 annotators was moderate, with a weighted Cohen kappa of 0.6. We used the average of their scores as the reference standard for evaluating the submitted systems. We then randomly selected 331 GE sentence pairs and 263 Epic sentences pairs. After combining these with the previous year's data set and removing duplicates, we finally obtained 1642 sentence pairs and released these as training data to each team to develop and fine-tune their systems. We used a total of 412 sentence pairs as the testing data set, including 189 GE sentence pairs (45.9%) and 223 Epic sentence pairs (54.1%), and asked the participating teams to return a numerical score indicating the degree of semantic similarity for each sentence pair.

Figure 1. Flowchart of the released data set generation in the 2019 n2c2/OHNLP track on Clinical Semantic Textual Similarity. EHR: electronic health record; PHI: protected health information.



Participating Teams

Participating teams were required to sign a Data Use Agreement to get access to the challenge data set. Each team could submit up to 3 runs for the testing data, with every run having 1 line for each sentence pair that provided the similarity score assigned by the system as a floating-point number.

Evaluation Metric

Similar to the general STS shared tasks, ClinicalSTS used the Pearson correlation coefficient between the predicted scores and the reference standard on the testing set to evaluate the submitted systems. We released a public script computing the Pearson correlation coefficient to the participating teams.

Results

Participating Teams

Figure 2 shows the number of teams that signed up the task, teams that submitted systems, and the total number of valid systems (ie, those outputs following the submission guideline), in comparison with the 2018 BioCreative/OHNLP ClinicalSTS shared task. In summary, 78 teams from 16 countries signed up for this shared task and 33 teams submitted a total of 87 valid

systems. Compared with the shared task in the previous year, the numbers of participating teams and submitted systems increased dramatically. Table 1 lists the details of teams that submitted systems, including team names, affiliations, and number of submitted systems.

Figure 2. Participation in the 2019 n2c2/OHNLP Clinical Semantic Textual Similarity (ClinicalSTS) track in comparison with the 2018 BioCreative/OHNLP Clinical STS track.

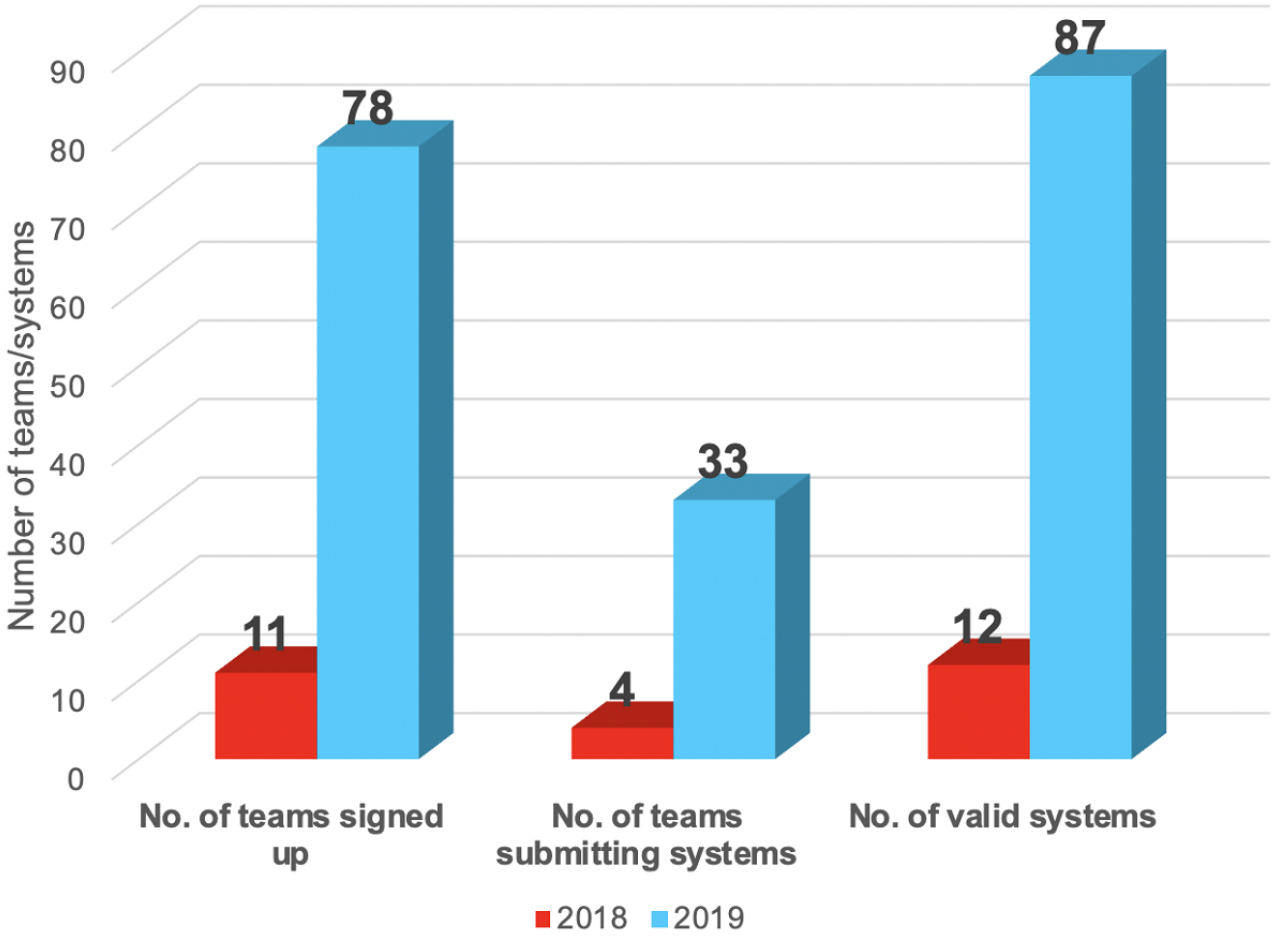


Table 1. Participating teams, affiliations, and number of systems submitted by each.

Team name	Affiliation	Number of systems
ASU	Arizona State University, USA	3
ChangYC	National Yang-Ming University, Taiwan	3
CLEARTeamCNRSLille	N/A[a]	3
DMSS	Boston Children's Hospital and Harvard University, USA	3
DUTIR	Dalian University of Technology, China	2
edmondzhang	Orion Health, USA	3
ezDI	ezDI Inc, USA	4
HITSZ	Harbin Institute of Technology at Shenzhen, China	3
IBMResearch	IBM Corporation, USA	0
JHU	Johns Hopkins University, USA	4
LSI_UNED	Universidad Rey Juan Carlos, Spain	3
MAH	Arizona State University, USA	3
MedDataQuest	Med Data Quest, USA	3
MICNLP	German Cancer Research Center, Germany	3
naist_sociocom	Nara Institute of Science and Technology, Japan	3
NCBI	National Center for Biotechnology Information, USA	3
nlpatvcu	Virginia Commonwealth University, George Mason University, USA	3
PUCPR	Pontifical Catholic University of Paraná, Brazil	2
QUB	Queen's University, UK	4
SBUnlp	Stony Brook University, USA	3
superficialintelligence0405	N/A	3
UAveiro	University of Aveiro, Portugal	3
UFL	University of Florida, USA	3
UH_RiTUAL	University of Texas at Houston, USA	3
Utah-VA	University of Utah and Veterans Affairs, USA	3
vjaneja	University of Maryland, USA	1
WSU-MQ	Western Sydney University, Australia	3
Yale	Yale University, USA	3
Yuxia	University of Melbourne, Australia	2
zhouxb	Yunnan University, China	3

[a]N/A: not available.

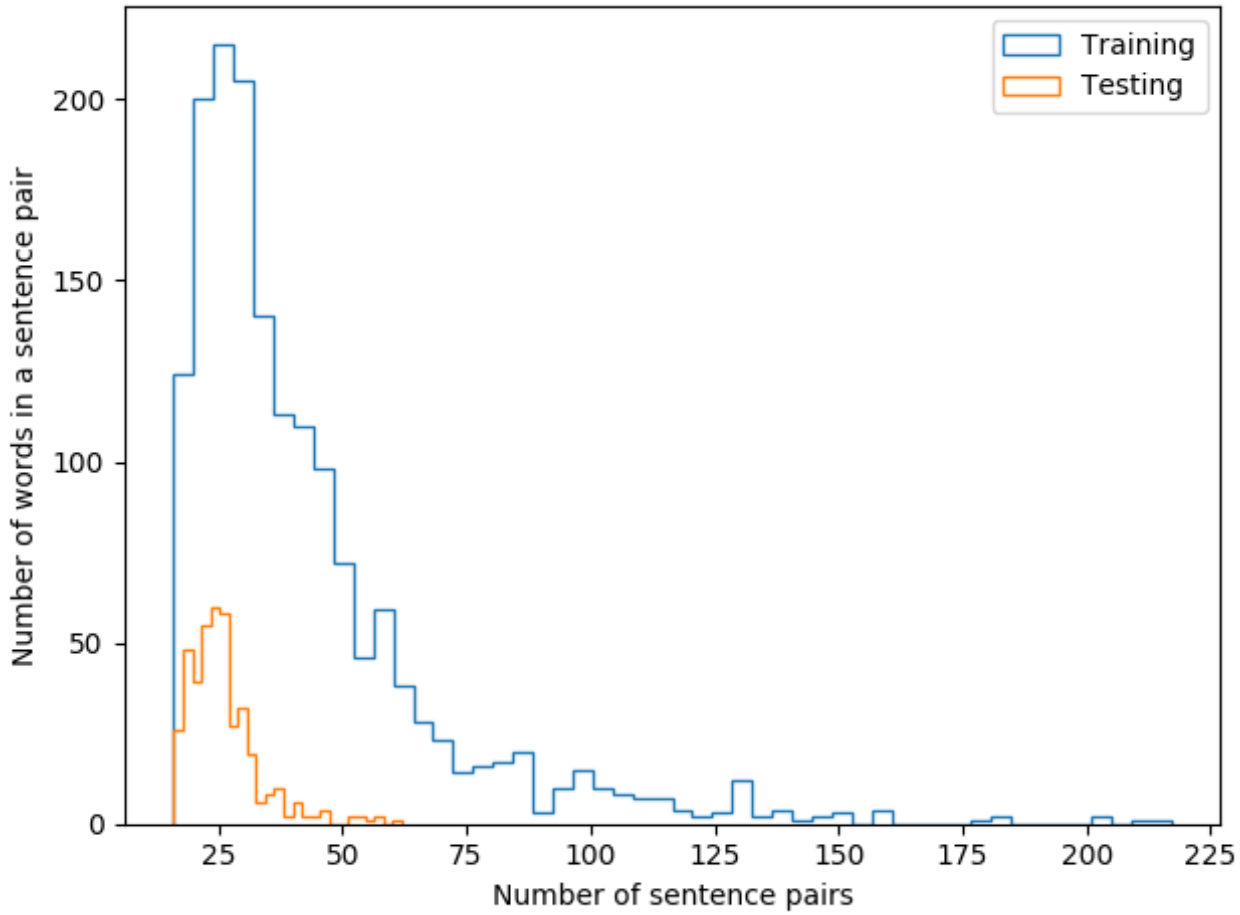
Basic Information of the Released Data Set

Our previous publication [13] provides more detailed information about the larger MedSTS data set. We used the Python NLP package spaCy version 2.1 (ExplosionAI GmbH) tokenizer to count the total number of words in each sentence pair. Figure 3 depicts the distribution of the number of words in sentence pairs in the released training and testing data sets. Most sentences pairs had 25 to 50 words, and there were more lengthy sentences in the training data set. However, the length distribution between training and testing was consistent. Table 2 lists the number of sentence pairs with different similarity scores in the released training and testing data sets. There were more sentence pairs with similarity scores between 3 and 4 in the training data set, whereas there were more sentence pairs with similarity scores between 1 and 2 in the training data set. This might have been due to sampling bias during data set creation.

Table 2. Number of sentence pairs with different similarity scores in the released training and testing data sets.

Similarity score	Training data set, n	Testing data set, n
[0,1)	185	98
[1,2)	236	168
[2,3)	245	30
[3,4)	607	34
[4,5]	369	82

Figure 3. Distribution of number of words in sentence pairs in the released training and testing data sets.



In addition, we used 3 surface lexical similarity methods as the baseline method to calculate the similarity scores in the testing data set, namely the Ratcliff/Obershelp pattern-matching algorithm, cosine similarity, and Levenshtein distance similarity. For more details about these baselines, please refer to our previous publication [13]. The cosine similarity achieved the best performance among the 3 baselines, with a Pearson correlation of r=.3709, followed by Levenshtein distance similarity with r=.2816 and Ratcliff/Obershelp with r=.2480.

Participating System Performance and Rankings

Table 3 lists the overall performance of all the valid submitted systems and the comparison with overall performance in the previous year's challenge. Table 4 shows the top 10 teams with their specific corresponding best runs and performance. The best system was from the team IBM Research's LM-POSTPROCESS-RUN with a Pearson correlation coefficient of r=.9010, an 8.2% increase from the previous year's best system. Overall, the median correlation score for the testing data set was r=.8291, a 3.4% increase from the previous year. We also compared the best run with other top systems using the Wilcoxon signed rank t test (Table 4). We found no statistically significant difference in 9 out of the top 10 systems (P<.001).

Table 3. Overall performance of the valid submitted systems and comparison with the previous year's results.

Metric	2019 n2c2/OHNLP ClinicalSTS[a], *r*	2018 BioCreative/OHNLP ClinicalSTS, *r*
Maximum	.9010	.8328
Minimum	–.0530	.7005
Median	.8291	.8016
Mean	.7183	.7820
Standard deviation	.2260	.0476

[a]ClinicalSTS: Clinical Semantic Textual Similarity.

Table 4. Performance of the top 10 teams with the corresponding best runs.

Rank	Team	Run	*r*	*P* value
1	IBMResearch	LM-POSTPROCESS-RUN	.9010	—[a]
2	NCBI	1	.8967	.88
3	UFL	XLNet-Run	.8864	.40
4	DMSS	AVERAGE-Run	.8792	.45
5	Yale	3	.8784	.09
6	QUB	fine_tuned_models_mean-Run	.8704	.54
7	MICNLP	Step1	.8694	<.001
8	HITSZ	raw_ensemble	.8685	.80
9	SBUnlp	ensembleall	.8677	.003
10	JHU	BERT-w-stsb-run	.8543	.005

[a]Not applicable.

We also compared the performance of valid systems for sentence pairs from GE and Epic EHR systems in the testing data set (Table 5). Overall, the participating systems performed better on the Epic sentence pairs than on the GE sentence pairs, despite the fact that a much larger portion of the training data were GE sentence pairs. This result indicates that the clinical sentences in our data set collected from the Epic EHR might be semantically simpler than those collected from the GE EHR system, which makes it easier for machine or deep learning models to learn the sentence semantic meaning.

Table 5. Performance comparison (Pearson correlation coefficient) between the Epic and GE sentence pairs.

Metric	Epic (n=223), *r*	GE (n=189), *r*
Maximum	.9148	.9022
Minimum	.0917	.0070
Median	.8377	.7785
Mean	.7792	.6812
Standard deviation	.1649	.2257

Table 6 shows the top 5 systems for the Epic and GE sentence pairs. The system from IBM Research achieved the best performance for the GE sentence pairs, which is consistent with their overall performance. Yale University's system (Run 4) had the best performance for the Epic sentence pairs, while the same system was not even in the top 5 performing systems for the GE sentence pairs.

Table 6. Top 5 systems for sentence pairs from the Epic and GE electronic health record systems.

Rank	Team	Run	*r*
Epic			
1	Yale	4	.9148
2	IBMResearch	LM-POSTPROCESS-RUN	.9098
3	NCBI	1	.9020
4	DMSS	AVERAGE-Run	.8949
5	UFL	Assemble-Run	.8863
GE			
1	IBMResearch	LM-POSTPROCESS-RUN	.9022
2	UFL	XLNet-Run	.9010
3	NCBI	1	.8938
4	Yale	3	.8796
5	MICNLP	Step1	.8576

Methods Used in the Participating Systems

Table 7 briefly summarizes the techniques used by the top teams. Most teams used state-of-the-art NLP neural language models in their systems, such as BERT [18] and XLNet [19], and state-of-the-art training schemas in deep learning, such as pretraining and fine-tuning schema, and multitask learning [20]. The outcomes from the top performing systems showed the advantages of these techniques over conventional machine learning and language models in learning semantics in human language, particularly in clinical language. Having said that, given the nature of the semantic simplicity of the sentences in the ClinicalSTS data set, neural language models and these training schemas need further comprehensive evaluation on larger clinical corpora with more complex sentences and semantics.

Table 7. Brief summary of the techniques used in the top systems.

Team	Techniques
IBMResearch	Multitask learning, BioBERT, RoBert, ClinicalBERT
NCBI	Convolutional neural network, multitask learning, BERT
UFL	BERT, XLNet
DMSS	BERT, XLNet
Yale	BERT, graph convolutional neural network
QUB	BERT, XLNet
MICNLP	BERT, medication graph
HITSZ	BERT, cTAKES
SBUnlp	BERT, Unified Medical Language System
JHU	BERT
Utah-VA	Multiple natural language processing features, deep neural network

Discussion

Principal Findings

We have given an overview of the 2019 n2c2/OHNLP ClinicalSTS shared task that aimed to measure the degree of semantic equivalence between 2 snippets of clinical text. We described how we prepared the data set in this year's shared task differently from that in the previous year, the participating teams and their systems, and the results. We witnessed an increasing research interest in the ClinicalSTS task among the NLP and medical informatics communities and increased system performance for the task. We also observed several limitations during the data preparation. There were limitations in the reference standard data creation, particularly for annotating the medication-related sentence pairs in the data set. Concerns were raised by participating teams regarding the judgement for those pairs. Table 8 shows an example of such a sentence pair. One may question that why the minocycline-oxycodone pair should have a much higher score than the oxycodone-pantoprazole pair. Minocycline is an antibiotic, and pantoprazole is an antacid. One annotator mentioned that the score of oxycodone + antibiotics was greater than the oxycodone + antacid score based on his experience of seeing them more frequently in the EHRs. In addition, the first case mentioned taking minocycline daily,

whereas the second case did not mention that pantoprazole should be taken once daily (such semantic information is missing in this case). Two of the annotators were nurses with a medical background but were not pharmacists. Both annotators agreed that in future work, involving pharmacists to annotate drug sentences could help make the annotation more accurate because drug sentences should be scored based on drug mechanisms, indications, doses, application period, and disease stages, plus pharmacogenomics and epigenomics or proteomics, etc.

Table 8. Examples of medication-related sentence pairs in the data set.

Examples	Score
sentence1: minocycline [MINOCIN] 100 mg capsule 1 capsule by mouth one time daily. sentence2: oxycodone [ROXICODONE] 5 mg tablet 1-2 tablets by mouth every 4 hours as needed.	3.0
sentence1: oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by mouth every 4 hours as needed. sentence2: pantoprazole [PROTONIX] 40 mg tablet enteric coated 1 tablet by mouth Bid before meals.	1.0

We also found that some sentence pairs seemed to be semantically equivalent but were assigned low similarity scores. For example, sentence 1 is "Thank you for choosing the Name, APRN, C.N.P., M.S. care team for your health care needs!" and sentence 2 is "Thank you for choosing the Name, M.D. care team for your health care needs!" The reason for the score (4.0) is that the degree of the provider is different. The provider in the first sentence is a nurse, whereas that in the second sentence is a physician. Thus, these 2 sentences are not equivalent. Another example is sentence 1: "Thank you for choosing the Name M.D. care team for your health care needs!" and sentence 2: "Thank you for allowing us to assist in the care of your patient." The reason for the score (2.0) is that the first sentence contains more details about the provider, whereas the second has fewer details.

Although there was a record number of 87 valid systems participating in the shared task, this is still not large enough to be able to extrapolate statistical analysis results to draw a convincing conclusion. The performance difference of these participating systems in the sentence pairs from different EHR systems may be attributable to bias in the system and the sampling data set.

In our future work, we might subcategorize the sentence pairs into different topics, such as medication or clinical workflow. We could provide tailored annotation guidelines according to the topic and invite subdomain experts with specific background (eg, pharmacist) to review sentences pairs in different topics (eg, medication-related sentence pairs).

Conclusions

ClinicalSTS is an important technique in many downstream clinical applications, such as clinical text summarization, clinical semantic information retrieval, and clinical decision support systems. In this paper, we provided an overview of the 2019 n2c2/OHNLP ClinicalSTS shared task that focused on computing semantic similarity for clinical text sentences generated from clinical notes in the real world. For this shared task, 33 international teams submitted a total of 87 valid systems. The top performing systems applied state-of-the-art NLP neural language models, such as BERT and XLNet, and state-of-the-art training schemas in deep learning, such as pretraining and fine-tuning schema. The best system used multitask learning and achieved a Pearson correlation coefficient of r=.9010, an 8.2% increase from the previous year's best system. We also compared the performance for sentences from both GE and Epic EHR systems and found better performance on the Epic sentence pairs than on the GE sentence pairs. The ClinicalSTS task remains challenging given the complexity of clinical texts. The ClinicalSTS shared task could continue to serve as a venue for researchers in NLP and medical informatics communities to develop and improve STS techniques for clinical text.

Acknowledgments

We would like to thank Donna Ihrke, RN, and Gang Liu, MD, for annotating the corpus. This work was made possible by funding support from the US National Center for Advancing Translational Sciences U01TR02062. We also would like to thank nference, Inc, and Linguamatics for their support for the 2019 n2c2/OHNLP workshop.

Conflicts of Interest

None declared.

References

1. Blumenthal D. Implementation of the federal health information technology initiative. N Engl J Med 2011 Dec 22;365(25):2426-2431 [FREE Full text] [doi: 10.1056/NEJMsr1112158] [Medline: 22187990]
2. Zhang R, Pakhomov S, McInnes BT, Melton GB. Evaluating measures of redundancy in clinical texts. AMIA Annu Symp Proc 2011;2011:1612-1620 [FREE Full text] [Medline: 22195227]
3. Wang MD, Khanna R, Najafi N. Characterizing the source of text in electronic health record progress notes. JAMA Intern Med 2017 Aug 01;177(8):1212-1213 [FREE Full text] [doi: 10.1001/jamainternmed.2017.1548] [Medline: 28558106]

4. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
5. Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. *SEM 2013 shared taskmantic textual similarity. In: Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. Stroudsburg, PA: Association for Computational Linguistics; 2013 Presented at: Second Joint Conference on Lexical and Computational Semantics (*SEM); Jun 13-14, 2013; Atlanta, GA, USA p. 32-43. [doi: 10.18653/v1/s17-2001]
6. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. Semeval-2014 task 10: multilingual semantic textual similarity. 2014 Presented at: 8th International Workshop on Semantic Evaluation (SemEval 2014); Aug 23-24, 2014; Dublin, Ireland p. 81-91. [doi: 10.3115/v1/s14-2010]
7. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. Semeval-2015 task 2: smantic textual similarity, English, Spanish and pilot on interpretability. 2015 Presented at: 9th International Workshop on Semantic Evaluation (SemEval); Jun 4-5, 2015; Denver, CO, USA p. 252-263. [doi: 10.18653/v1/s15-2045]
8. Agirre E, Banea C, Cer D, Diab M, Gonzalez-Agirre A, Mihalcea R, et al. Semeval-2016 task 1-smantic textual similarity, monolingual and cross-lingual evaluation. 2016 Presented at: 10th International Workshop on Semantic Evaluation (SemEval-2016); Jun 16-17, 2016; San Diego, CA, USA p. 497-511. [doi: 10.18653/v1/s16-1081]
9. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation. 2017 Presented at: 11th International Workshop on Semantic Evaluation (SemEval-2017); Aug 3-4, 2017; Vancouver, BC, Canada p. 170800055 URL: https://www.aclweb.org/anthology/S17-2001.pdf [doi: 10.18653/v1/s17-2001]
10. Afzal N, Wang Y, Liu H. MayoNLP at SemEval-2016 task 1: semantic textual similarity based on lexical semantic net and deep learning semantic model. 2016 Presented at: 10th International Workshop on Semantic Evaluation (SemEval-2016); Jun 16-17, 2016; San Diego, CA, USA p. 674-679. [doi: 10.18653/v1/s16-1103]
11. National NLP Clinical Challenges (n2c2). 2019 n2c2 shared-task and workshop. Track 1: n2c2/OHNLP track on clinical semantic textual similarity. Boston, MA: Harvard Medical School URL: https://n2c2.dbmi.hms.harvard.edu/track1 [accessed 2020-11-09]
12. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L, Shen F, et al. Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. 2018 Presented at: BioCreative/OHNLP Challenge; Aug 29-Sep 1, 2018; Washington, DC, USA. [doi: 10.1145/3233547.3233672]
13. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resources Eval 2018 Oct 24;54(1):57-72. [doi: 10.1007/s10579-018-9431-1]
14. Li D, Rastegar-Mojarad M, Elayavilli R, Wang Y, Mehrabi S, Yu Y, et al. A frequency-filtering strategy of obtaining PHI-free sentences from clinical data repository. New York, NY: Association for Computing Machinery; 2015 Presented at: 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics; Sep 9-12 2015; Atlanta, GA, USA p. 315-324. [doi: 10.1145/2808719.2808752]
15. Black PE. Ratcliff/Obershelp pattern recognition. In: Pieterse V, Black PE, editors. Dictionary of Algorithms and Data Structures. Gaithersburg, MD: National Institute of Standards and Technology; 2004. URL: https://www.darkridge.com/~jpr5/mirror/dads/HTML/ratcliffObershelp.html [accessed 2020-11-09]
16. 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 2017 Nov 30. [doi: 10.1093/jamia/ocx138] [Medline: 29202185]
17. Soukoreff RW, MacKenzie IS. Measuring errors in text entry tasks: an application of the Levenshtein string distance statistic. New York, NY: Association for Computing Machinery; 2001 Presented at: CHI'01 Extended Abstracts on Human Factors in Computing Systems; Mar 2001; Seattle, WA, USA p. 319-320. [doi: 10.1145/634067.634256]
18. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, MN, USA p. 1810 URL: https://arxiv.org/abs/1810.04805
19. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: generalized autoregressive pretraining for language understanding. Adv Neural Inf Processing Syst 2019;2019:5753-5763.
20. Ruder S. An overview of multi-task learning in deep neural networks. arXiv. 2017. URL: https://arxiv.org/abs/1706.05098 [accessed 2020-11-11]

Abbreviations

ClinicalSTS: Clinical Semantic Textual Similarity
EHR: electronic health record
n2c2: National NLP Clinical Challenges
NLP: natural language processing
OHNLP: Open Health Natural Language Processing
PHI: protected health information

STS: semantic textual similarity

Edited by C Lovis; submitted 10.08.20; peer-reviewed by J Du, J Rousseau, J Li; comments to author 06.09.20; revised version received 16.10.20; accepted 03.11.20; published 27.11.20.

Please cite as:
Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H
The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview
JMIR Med Inform 2020;8(11):e23375
URL: http://medinform.jmir.org/2020/11/e23375/
doi:10.2196/23375
PMID:33245291

Original Paper

Toward Preparing a Knowledge Base to Explore Potential Drugs and Biomedical Entities Related to COVID-19: Automated Computational Approach

Junaed Younus Khan[1], BSc; Md Tawkat Islam Khondaker[1], BSc; Iram Tazim Hoque[1], BSc; Hamada R H Al-Absi[2], MSc; Mohammad Saifur Rahman[1], PhD; Reto Guler[3,4,5], PhD; Tanvir Alam[2], PhD; M Sohel Rahman[1], PhD

[1]Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

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

[3]International Centre for Genetic Engineering and Biotechnology, Cape Town Component, Cape Town, South Africa

[4]Division of Immunology and South African Medical Research Council Immunology of Infectious Diseases, Department of Pathology, Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

[5]Wellcome Centre for Infectious Diseases Research in Africa, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

Corresponding Author:
Tanvir Alam, PhD
College of Science and Engineering
Hamad Bin Khalifa University
PO Box 34110
Education City
Doha
Qatar
Phone: 974 44542277
Email: talam@hbku.edu.qa

Abstract

Background: Novel coronavirus disease 2019 (COVID-19) is taking a huge toll on public health. Along with the non-therapeutic preventive measurements, scientific efforts are currently focused, mainly, on the development of vaccines and pharmacological treatment with existing drugs. Summarizing evidences from scientific literatures on the discovery of treatment plan of COVID-19 under a platform would help the scientific community to explore the opportunities in a systematic fashion.

Objective: The aim of this study is to explore the potential drugs and biomedical entities related to coronavirus related diseases, including COVID-19, that are mentioned on scientific literature through an automated computational approach.

Methods: We mined the information from publicly available scientific literature and related public resources. Six topic-specific dictionaries, including human genes, human miRNAs, diseases, Protein Databank, drugs, and drug side effects, were integrated to mine all scientific evidence related to COVID-19. We employed an automated literature mining and labeling system through a novel approach to measure the effectiveness of drugs against diseases based on natural language processing, sentiment analysis, and deep learning. We also applied the concept of cosine similarity to confidently infer the associations between diseases and genes.

Results: Based on the literature mining, we identified 1805 diseases, 2454 drugs, 1910 genes that are related to coronavirus related diseases including COVID-19. Integrating the extracted information, we developed the first knowledgebase platform dedicated to COVID-19, which highlights potential list of drugs and related biomedical entities. For COVID-19, we highlighted multiple case studies on existing drugs along with a confidence score for their applicability in the treatment plan. Based on our computational method, we found Remdesivir, Statins, Dexamethasone, and Ivermectin could be considered as potential effective drugs to improve clinical status and lower mortality in patients hospitalized with COVID-19. We also found that Hydroxychloroquine could not be considered as an effective drug for COVID-19. The resulting knowledgebase is made available as an open source tool, named COVID-19Base.

Conclusions: Proper investigation of the mined biomedical entities along with the identified interactions among those would help the research community to discover possible ways for the therapeutic treatment of COVID-19.

(JMIR Med Inform 2020;8(11):e21648) doi:10.2196/21648

KEYWORDS

COVID-19; 2019-nCoV; coronavirus; SARS-CoV-2; SARS; remdesivir; statin; statins; dexamethasone; ivermectin; hydroxychloroquine

Introduction

SARS-CoV-2 initially spread widely in China, then in Italy, and has since been reported worldwide [1,2]. SARS-CoV-2 is a novel coronavirus that causes COVID-19 [3]. Although SARS-CoV-2 has gained attention as a consequence of the global COVID-19 pandemic, other known human coronaviruses, including betacoronaviruses (SARS-CoV, MERS, OC43, HKU1) and alphacoronaviruses (229E, NL63), have resulted in severe respiratory syndrome in patients and been of public health concern [4]. To combat COVID-19, an urgent solution is needed for the detection and therapeutic treatment of this disease, which requires a comprehensive experimental investigation of relevant biomedical entities (eg, genes, noncoding ribonucleic acids [ncRNA], viruses, drugs) [5]. However, this is a relatively slow process due to the inherent nature of experimental validation. As an alternative, faster in silico methods can be applied [6,7], which can act as a filter prior to wet lab validation. Virtual screening, molecular docking, and other in silico methods have already been investigated to discover drugs that may work against COVID-19 [8]. Still, this is a daunting task due to the large number of possible combinations of biomedical entities (eg, drug-gene pairs) that need to be examined [9]. To enable comprehensive exploration of potential therapeutic treatments, knowledge base solutions are proposed; these would allow the scientific community to focus on a relatively smaller number of potential biomedical entities that may lead to the discovery of a novel treatment for COVID-19.

Databases that focus on virus-related diseases for multiple hosts already exist. For example, in ViRBase [10], the authors highlighted the association between ncRNAs and viruses in 20 hosts. The VISDB database, based on literature mining, integrated the virus interaction site in humans for five DNA oncoviruses and four RNA retroviruses [11]. Virus Pathogen Resources (VIPR) developed a portal that collected a comprehensive set of information related to coronavirus and hepatitis C virus (HCV), as well as other viruses [12,13]. However, none of the abovementioned databases are particularly useful for COVID-19/SARS-CoV-2, as those databases were not specific to the novel coronavirus, or they provided very limited information about the associated genes, or they did not include other factors involved in coronavirus-related diseases, drugs, and drug side effects. Moreover, there is no one knowledge base that has integrated all biomedical entities specific to COVID-19/SARS-CoV-2. To address this gap, we explored the potential of machine intelligence to automatically mine the scientific literature, with the goal of developing the first comprehensive knowledge base that integrates several biomedical entities associated with COVID-19/SARS-CoV-2. To achieve this, we leveraged state-of-the-art natural language processing algorithms, sentiment analysis, and deep learning–based techniques and applied them to a large corpus of coronavirus-related scientific literature.

Methods

Data Sets

For this study, we used the COVID-19 Open Research Dataset (CORD-19) [14], generated by the Allen Institute for AI. The data set contains over 138,000 scholarly articles related to COVID-19 and the coronavirus family of viruses. The data set was collected using the following query to search PubMed, PubMed Central (PMC), bioRxiv, and medRxiv: “COVID-19” OR “Coronavirus” OR “Corona virus” OR “2019-nCoV” OR “SARS-CoV” OR “MERS-CoV” OR “Severe Acute Respiratory Syndrome” OR “Middle East Respiratory Syndrome.” This query covers most research articles related to COVID-19 and other coronaviruses (eg, MERS, SARS) and we searched up until June 9, 2020. Unless otherwise specified, we considered both the abstract and full body of the manuscripts (when available) for downstream analysis.

Source of Dictionaries

We collected gene names from HUGO Gene Nomenclature Committee (HGNC) [15], Protein Data Bank (PDB) entries from PDB [16], micro ribonucleic acids (miRNAs) from miRBase [17], disease names from Disease Ontology (DO) [18], drug names from DrugBank [19], and drug side effects from Side Effect Resource (SIDER) [20].

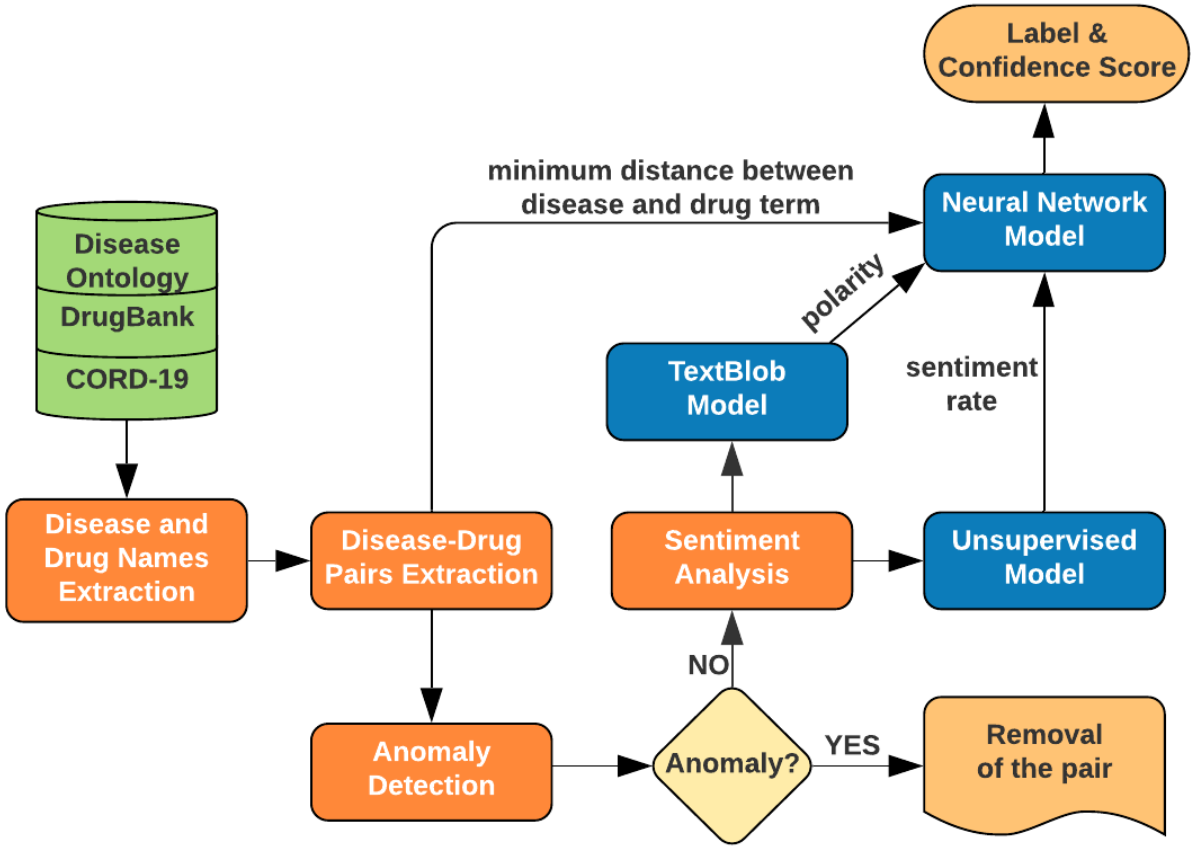
Overview of Methodology

We extracted disease-drug, disease-gene, drug-PDB pairs, and their corresponding sentences from the CORD-19 literature in a co-occurrence–based approach. To evaluate the effectiveness of the disease-drug pairs, we used both a pretrained model (TextBlob) and an unsupervised model (developed by the authors using the Word2Vec model and K-means clustering) to determine the sentiment scores of the sentences extracted for each pair. We further used these sentiment scores along with the minimum distance between the disease and drug term in the corresponding sentences as input features of our neural network model, which we used for the final classification of the disease-drug pairs (as positive or negative). To determine the confidence level of the extracted disease-gene associations, we transformed each disease and gene of a pair into two separate vectors using the Word2Vec model and calculated their cosine similarity. We used the known disease-gene associations from the DisGeNET database as the gold standard to determine the confidence level of the new associations on the basis of cosine similarity measures. Finally, we extracted the side effects of the drugs that were found by our mining from SIDER. Additionally, a feedback mechanism was incorporated into COVID-19Base to collect feedback from users for future use.

Extracting Disease-Drug Interactions

We extracted disease-drug interactions from the CORD-19 literature and classified them into one of two categories (labels): positive and negative. The positive label means the drug is potentially effective against COVID-19, and the negative label means the opposite. We also determined a confidence score, which indicates our level of confidence in that automatic label. Figure 1 shows the workflow of extracting disease-drug interactions and predicting the effectiveness of drugs against diseases with confidence scores.

Figure 1. Flowchart of extracting disease-drug interactions and predicting the effectiveness of drugs against diseases with confidence scores. CORD-19: COVID-19 Open Research Dataset.



Disease and Drug Name Extraction

To extract relevant disease-drug pairs from the CORD-19 literature, we employed a dictionary-based approach to detect mentions of diseases and drugs in the literature. We used Disease Ontology [18] and DrugBank [19] to prepare the disease and drug dictionaries. We leveraged the Aho-Corasick algorithm [21] to search the drug and disease names, considering the large size of the drug and disease dictionaries and the corpus itself. The Aho-Corasick algorithm is a string-searching algorithm that efficiently locates multiple patterns in a large amount of text. The time complexity of the algorithm is $O(n + m + z)$, where n is the length of the text, m is the total length of all the patterns to be searched, and z is the total number of occurrences of the patterns in the text.

Disease-Drug Pairs Extraction

After extracting the disease and drug names separately, we wanted to mine the literature and identify the sentences that contain the disease and drug pairs to semantically evaluate their interactions. For this purpose, we searched for every disease-drug pair from our disease and drug list in the CORD-19 literature and collected every sentence where a co-occurrence was found. We then created a document for every disease-drug pair, combining all extracted sentences. Thus, we built a disease-drug pair to document mapping. We did not use a pattern-based approach here (as was done previously in [22]) as this could result in missing some sentences containing disease-drug pairs.

Anomaly Removal

As we automatically extracted the sentences containing the disease-drug pairs, there was a possibility of errors in our extracted data; therefore, we decided to check and remove any abnormalities from our collected data before moving on to the next stage of the pipeline. We used unsupervised anomaly detection [23] for this task. Unsupervised anomaly detection detects anomalies in an unlabeled data set by looking for instances that seem to fit the remainder of the data set the least, under the assumption that the majority of the instances in the data set are “normal.” We used the K-means clustering algorithm

[24], as it has been used for anomaly detection in several studies [25-29]. We proceeded as follows. First, we used Doc2Vec [30] to create a numeric representation of each document associated with each disease-drug pair. We then fitted these representations into our K-means model and observed two clear clusters of easily discriminable sizes, where the smaller one consisted of only 189 instances. As we know that anomalies differ from the normal instances significantly and occur very rarely in the data, we could assume that the instances of the smaller cluster were indeed anomalies. We also checked a number of instances manually to verify our assumption. We discarded these 189 instances from any further consideration.

Sentiment Analysis

Overview

We applied sentiment analysis to automatically assess the effectiveness of a drug to treat a particular disease in the context of each extracted drug-disease pair. First, we applied the concept of transfer learning. We used TextBlob [31], which is a pretrained sentiment analysis tool provided as a Python library. However, it showed some inconsistency in some cases as expected from a pretrained model and we felt it necessary to perform unsupervised sentiment analysis, which is the second model in our pipeline. We obtained a polarity score from the TextBlob model and a sentiment rate from our unsupervised model for each disease-drug pair, which were subsequently fed to our neural network model to predict the final label.

TextBlob Model

TextBlob is a Python library that is widely used in natural language processing tasks such as part-of-speech (POS) tagging, noun phrase extraction, sentiment analysis, classification, and translation. Given the sentences that we mined for each disease-drug pair as input, TextBlob gives a polarity score between –1 and 1. We recorded the polarity scores for each disease-drug pair to use it as a feature for our neural network model.

Unsupervised Model

We used the concept of K-means clustering again for unsupervised sentiment analysis. First, we trained the Word2Vec [32] model with our mined literature and got a vector representation of every word. We then ran K-means clustering on the estimated word vectors and found two clusters (positive and negative). The positive cluster was decided on the basis of the presence of several positive words (in the context of a disease-drug pair), including “cure,” “preclude,” “inhibit,” “prescribe,” “reduce,” and “modest.” On the other hand, the negative cluster contained words like “risky,” “kill,” and “danger.” We then assigned each word a sentiment value, either +1 or –1, based on the cluster (positive or negative) they belong to. We weighed this value by dividing it by the distance between the word and the centroid of its cluster to describe the extent of its potential positiveness or negativeness. We then calculated the term frequency–inverse document frequency (tf–idf) score [33,34] of each word in the sentence collection to consider the significance of the unique words. Next, we built a tf–idf representation, T, for each disease-drug pair by replacing each word of the corresponding sentences with its tf–idf score and a sentiment value representation, S, by replacing each word with its sentiment value. Finally, we took their dot product ($T \cdot S$) as the final sentiment rate of our unsupervised model.

Neural Network Model for Automatic Labeling and Confidence Score

Overview

We used a deep neural network (DNN) model to automatically predict the label and confidence score for our disease-drug pairs. We used a relatively simpler neural network with two hidden layers as such models commonly perform better for smaller data sets compared to neural networks with many layers and parameters [35,36].

Training Data

We manually labeled 200 disease-drug pairs to train our neural network model. Among them, there were 110 positive instances and the rest were negative.

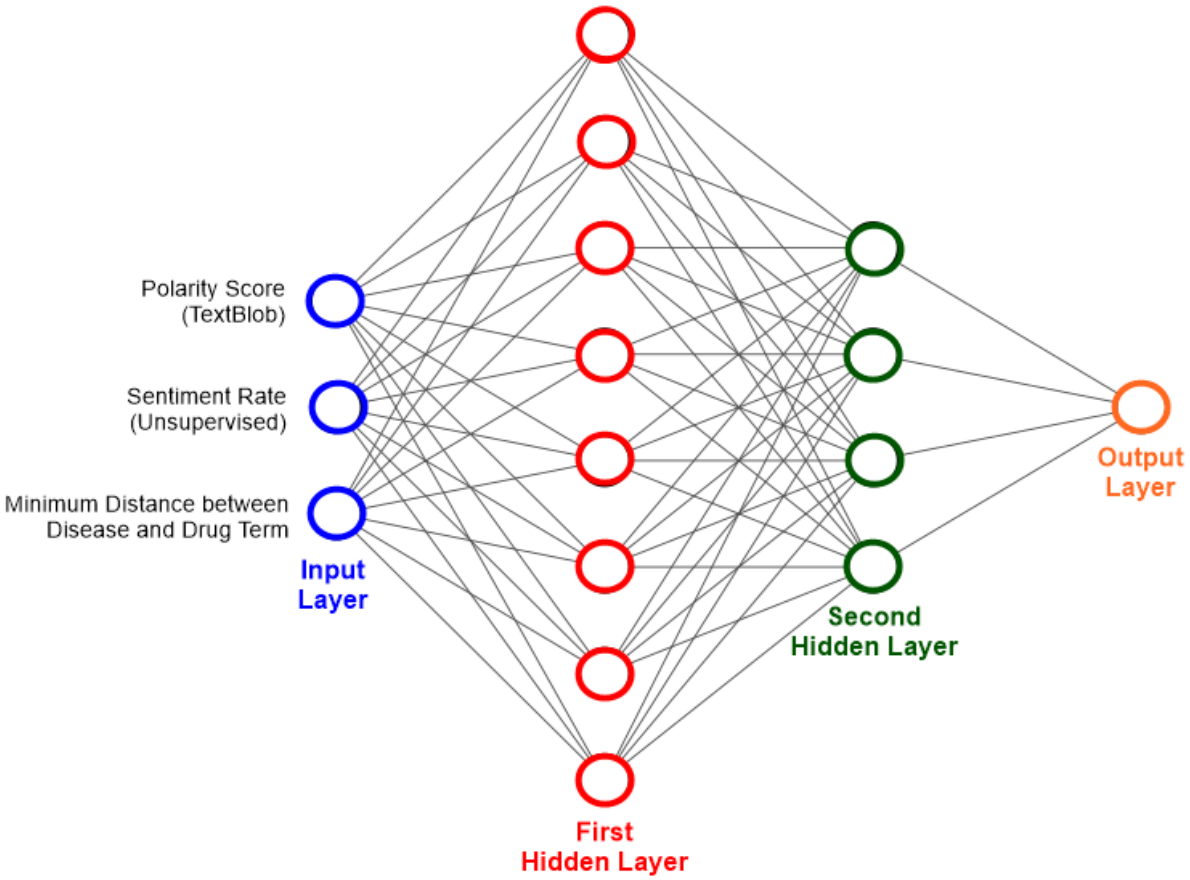
Input Features

We used the polarity or sentiment score given by the TextBlob and unsupervised models as the input features for our neural network model, along with the minimum distance between the disease and drug term in the corresponding document.

Model Setup and Output

The DNN structure used in this study is similar to that shown in Figure 2. It consists of one input layer with three neurons (each neuron corresponds to one input feature), two hidden layers with eight and four neurons respectively, and one output layer containing one neuron for binary classification (positive or negative). The transfer functions of the first and second hidden layers were the rectified linear unit (ReLU) [37] and hyperbolic tangent function (tanh) [38], respectively. The transfer function of the output layer was a sigmoid function [39]. We trained the DNN model using Xavier initialization [40], which tries to make the variance of the outputs of a layer equal to the variance of its inputs. We used Adam optimizer [41] and the maximum training epoch was set to 500. We split our labeled data into training and test sets on an 80:20 ratio. We trained our model on the training data and achieved 75% accuracy on the test set.

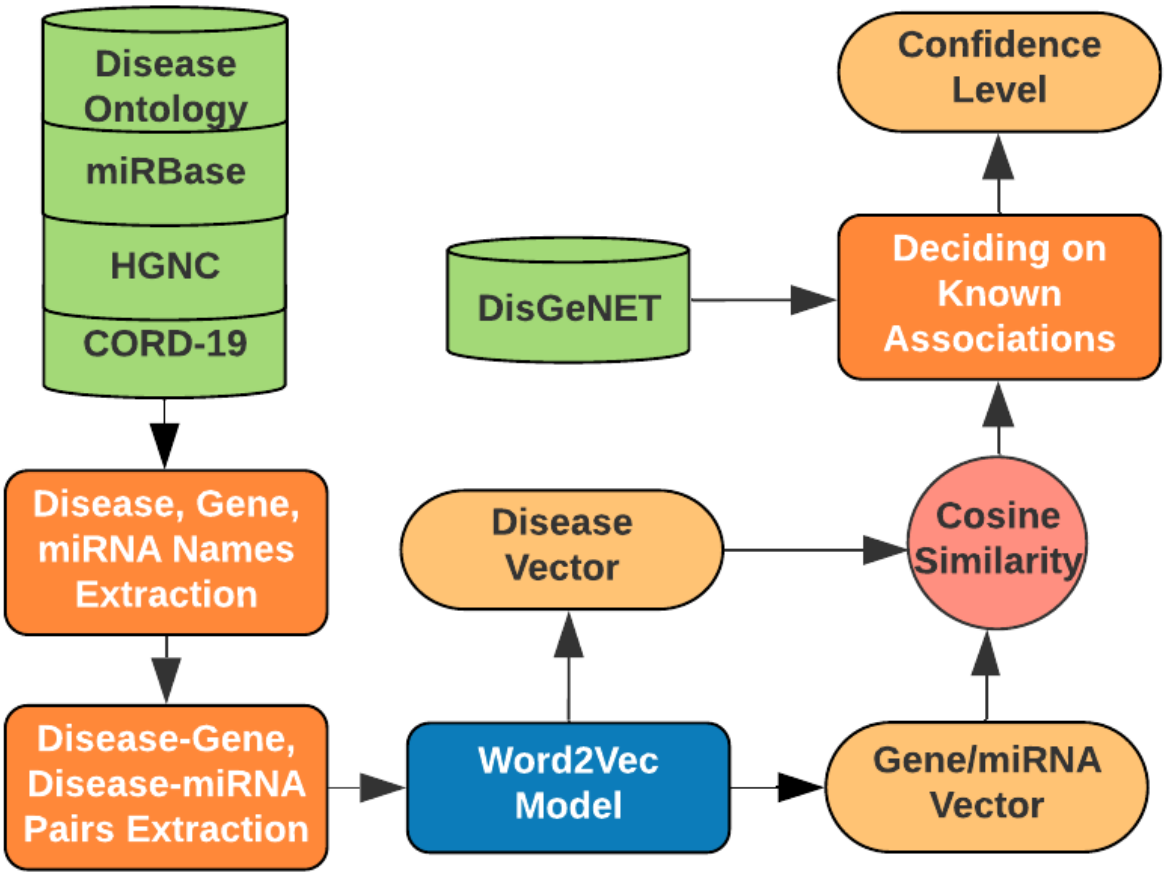
Figure 2. Schematic diagram of the deep neural network used to predict the effectiveness of drugs against diseases.



Extracting Disease-Gene Associations

Figure 3 shows the workflow of extracting disease-gene associations. We extracted gene names along with miRNAs from the CORD-19 literature in a dictionary-based approach using HGNC [15] and miRBase [17]. We then extracted their associations with diseases in a similar process to the one we had used to extract the disease-drug pairs and collected all the abstracts where a co-occurrence was found. Next, we applied the concept of cosine similarity [42] to confidently infer the associations. We transformed each disease into vector V_1, each gene (and miRNA) into vector V_2, and then calculated the cosine similarity of V_1 and V_2 for each pair. To create the vector representations, we trained a Word2Vec model with all the collected abstracts. We used the DisGeNET [43] database as the gold standard to evaluate the performance of cosine similarity in predicting the gene-disease linkage. First, we calculated the maximum, average, and minimum cosine similarity of the pairs that were common both in our findings and in the DisGeNET database. We found that 99.7% of the newly discovered pairs lie within this range (determined from DisGeNET) in terms of cosine similarity. We further classified the associations into three classes (high, medium, and low) in terms of confidence as follows: pairs having cosine similarity closest to the maximum (minimum) of the known ones were considered as high (low) confidence associations, and the remaining ones (those closest to the average) as medium confidence associations. Moreover, pairs that were also found in the DisGeNET database were labeled as verified associations.

Figure 3. Flowchart of extracting disease-gene and disease-miRNA associations and determining their confidence levels. CORD-19: COVID-19 Open Research Dataset; HGNC: HUGO Gene Nomenclature Committee; miRNA: micro ribonucleic acid.



Extracting Drug-Protein Associations

We also extracted drug-protein associations from the CORD-19 literature, applying the same co-occurrence–based approach as mentioned above. We used PDB IDs from the Protein Data Bank [16] for extracting protein names. Unlike the disease-gene associations, we did not apply the concept of cosine similarity here as we did not find any suitable data set that could be used as the gold standard in this case.

Extracting Side Effects of Drugs

The drugs we are suggesting through this literature mining may come with different side effects. Therefore, we also explored the possible side effects of the drugs. We collected the drugs with their corresponding side effects from SIDER [20] and mapped them with the drugs mentioned in the CORD-19 literature to extract the possible side effects.

Feedback Mechanism

We implemented a feedback mechanism in COVID-19Base for future improvement. This mechanism enables expert users from the scientific community to share their valuable feedback on the label (positive or negative) for a particular interaction determined by the automatic natural language processing–based approach. The users can voluntarily label each sentence that is mined from the literature as a source of an interaction. This feedback will be recorded and further processed to enrich the labeled data set, which can be leveraged in the next version of COVID-19Base to further improve the prediction quality for determining effective disease-drug interactions. The accompanying tutorial (user manual) on COVID-19Base highlighted an example of how a user can use the feedback mechanism.

Results

Terms and Interactions Highlighted in CORD-19 Data Set

Based on our computational workflow, we identified 1805 diseases, 2454 drugs, 1910 genes, 11 miRNAs, and 70 PDB entries from the CORD-19 literature (Table 1). Among the disease-drug pairs, 21,581 were positive and 1318 were negative. Among the disease-gene associations, 2088 were verified, and 82 associations were found with high-confidence, 12,231 with medium-confidence, and 1488 with low-confidence. More results are shown in Table 1. Notably, a small proportion (1.5%) of the findings were manually labeled. Interestingly, we found 194 drug-PDB pairs for coronavirus-related diseases, which indicates the rapid growth of experimental work to understand the interaction mechanisms of drugs and target proteins.

Table 1. Pairs of terms as identified in the analyzed set of documents[a].

Interaction or association	Number of extracted pairs of terms
Disease-drug	22,899 (21,581 positive, 1318 negative)
Disease-gene	15,889 (2088 verified, 82 high, 12,231 medium, 1488 low)
Disease-miRNA	56 (48 medium, 8 low)
Drug-Protein Data Bank	194

[a]Positive (negative) indicates an (in)effective association. High, medium, and low refer to confidence associations.

COVID-19–Related Terms and Interactions

Our computational workflow identified 514 drugs and 417 genes that are directly associated with COVID-19 (Table 2). Among the 514 drugs, 492 were found to have a positive association and 22 had a negative association. Among the 417 genes, 347 were medium-confidence associations and 70 were low-confidence associations.

Table 2. Biomedical terms that are related to COVID-19[a].

Interaction or association	Number of extracted pairs of terms
COVID-19–drug	514 (492 positive, 22 negative)
COVID-19–gene	417 (347 medium, 70 low)
COVID-19–miRNA	3 (2 medium, 1 low)

[a]Positive (negative) indicates an (in)effective association. High, medium, and low refer to confidence associations.

Genes Related to COVID-19

Our automated workflow identified C-reactive protein (CRP) as one of the COVID-19–associated genes with "medium" confidence. CRP is a known clinical biomarker for SARS [44] and the level of CRP increases significantly in patients with SARS. The level of CRP was also higher for patients with COVID-19 in some clinical cases [45,46]. More than 25 papers (from the CORD-19 data set) related to the association between CRP and COVID-19 were identified through our computational workflow. Furthermore, the genes *ELANE*, *AZU1*, *MPO*, *PRTN3*, *CTSG*, and *TCN1* were shown to be significantly altered in patients with COVID-19 [47], and our automatically prepared knowledge base highlights all of them as associated with COVID-19 with "medium" or "low" confidence. The *ACE2* and *TMPRSS2* genes are known to be involved in SARS-CoV-2 infection [48]; in fact, SARS-CoV-2 uses angiotensin-converting enzyme 2 (ACE2) as a receptor for entry into host cells [49,50]. The spike protein of SARS-CoV-2 binds with the ACE2 receptor and the protease TMPRSS2 mediates the infection process [51]. It is important to note that *ACE2* and *TMPRSS2* were not directly listed in DisGeNET as genes associated with COVID-19. In spite of that, our data-driven approach based on a gold-standard data set from DisGeNET was able to infer the association of *ACE2* and *TMPRSS2* with COVID-19 with "medium" confidence, which suggests that our approach is efficacious. Analyzing the complete *ACE2* interaction network, Wicik et al [52] listed several element genes (*ACE2*, *ANPEP*, *DPP4*, *CCL2*, *MEPIA*, *TFRC*, *ADAM17*, *NPC1*, *FABP2*, *TMPRSS2*, *CLEC4M*) and all of these genes were identified as COVID-19–associated in our automatically prepared knowledge base. In addition, we mined three miRNAs (hsa-miR-4661-3p, hsa-miR-429, and hsa-miR-183) that were mentioned in the abstracts of COVID-19–related literature.

Case Studies

In this section, we discuss interesting and useful findings from our automatically prepared knowledge base in the context of potential drugs that can be investigated for the potential therapeutic treatment of COVID-19.

Case Study 1: Dexamethasone Can be Considered an Effective Drug for COVID-19

Dexamethasone, an inexpensive and commonly used steroid, is a major breakthrough in the fight against COVID-19. We found dexamethasone to be a positive (ie, effective) drug for COVID-19, automatically labeled as such through our computational workflow with a confidence score of 77.61%. Our computational workflow also discovered the effectiveness of this drug against pneumonia, respiratory failure, and diarrhea, which are strongly correlated to COVID-19 [53,54]. Thus, further exploration of this drug to fight COVID-19 is likely to be fruitful. Recent studies suggest that dexamethasone reduces the risk of death from COVID-19 from 40% to 28% for patients on ventilators and from 25% to 20% for patients needing oxygen [55].

Case Study 2: Ivermectin Might be Considered an Effective Drug for COVID-19

Ivermectin is an effective drug against pneumonia and diarrhea, and has recently been claimed to successfully treat patients with COVID-19 as well [56]. It is a US Food and Drug Administration (FDA)–approved drug used for parasitic infections, which has the potential to be repurposed. Ivermectin inhibits the replication of SARS-CoV-2 in vitro [57]. Recently, a team of medical doctors in Bangladesh reported quick recoveries of patients with COVID-19 using this drug [58]. We found ivermectin to be a positive (ie, effective) drug for COVID-19, automatically labeled with a confidence score of

77.91%. It was also labeled a positive drug for pneumonia and diarrhea in our knowledge base.

Case Study 3: Remdesivir Seems Effective Against COVID-19

Remdesivir has been identified as a positive (ie, effective) drug for COVID-19, automatically labeled as such through our pipeline, with a confidence score of 68.18%. Thus, it seems to be a promising drug for further investigation for treating COVID-19. Interestingly, it was recently being considered as an effective drug for treating COVID-19 [59]. Notably, remdesivir is an antiviral drug originally developed for Ebola treatment [60,61]. A recent clinical trial conducted by the National Institute of Allergy and Infectious Diseases (NIAID) showed that remdesivir helped patients with COVID-19 recover faster and improved their survival rates. Adult patients treated with remdesivir were found to recover 4 days faster, an improvement of 31% compared to other patients; in addition, the overall death rate dropped from 11.6% to 8% [62]. Remdesivir is now under consideration for use against COVID-19 in more than ten clinical trials [63]. We found 6LU7 was one of the PDB entries for remdesivir. After exploring the corresponding literature [64], we found that remdesivir was shown to be an effective inhibitor of the main SARS-CoV-2 protease using molecular docking [65,66].

Case Study 4: Hydroxychloroquine Is Not an Effective Treatment for COVID-19

Antimalaria drug hydroxychloroquine, which is one of the most talked-about drugs for treating COVID-19, was also found in our mining, albeit with a negative interaction. Our model found it is a negative (ie, ineffective) drug with 64.67% confidence. Additionally, it also revealed that this drug has 111 side effects including anemia, hemorrhage, liver disorder, hepatitis fulminant, cardiomyopathy, and cardiac failure, which makes it a risky option, especially for patients with heart and liver complications. Although the FDA had previously granted authorization to use this drug for COVID-19, it has recently cautioned against its use outside of a hospital setting or a clinical trial due to its side effects and risk factors [67].

Case Study 5: Statins Drugs Could be Effective Against COVID-19

Statins are effective as lipid-lowering drugs and mainly used for the treatment of cardiovascular diseases [68]. Statins are also well known for their anti - inflammatory effects [69] and some studies have supported the use of these drugs as part of a COVID - 19 treatment protocol [70]. Multiple clinical trials (eg, NCT04343001, NCT04380402) have been launched to determine the efficacy of statins against COVID-19 [71,72]. In our knowledge base, the majority of statin classes were shown to be effective against COVID-19. For example, ulinastatin, rosuvastatin, fluvastatin, and lovastatin were labeled as positive (ie, effective) drugs against COVID-19 with 94.04%, 79.38%, 78.88%, and 70.75% confidence scores, respectively. Through our automated computational workflow, we found only one mention of atorvastatin in the literature [73]. In that single article, Deliwala et al [73] mentioned atorvastatin as part of a prevention plan against cortical stroke for a 31-year-old female patient with COVID-19, without referring to the effectiveness of atorvastatin against COVID-19. Consequently, our knowledge base labeled atorvastatin with a negative sentiment and a rather low confidence score (61.22%) for COVID-19. We anticipate that as the number of articles related to atorvastatin use in COVID-19 treatment protocols increases, our model will be able to effectively infer the sentiment (effective versus ineffective) of this drug. Based on our finding, it is safe to state that statins, as low-cost and well-tolerated drugs, should be investigated in more detail in clinical trials; such drugs may help low- and middle-income countries in particular, where expensive drugs might not be affordable.

Discussion

Principal Findings

In our knowledge base, through a computational workflow, we not only extracted the drugs and other biomedical terms that are mentioned in the literature, but also identified “term pairs” based on their co-occurrence, which will allow the scientific community to investigate in depth the associations between term pairs like disease-gene and disease-drug. Many drugs were associated with COVID-19, representing the cumulative effort of the scientific community to repurpose existing drugs rather than pursue novel drug discoveries, which is a rational approach in a pandemic situation [74]. We leveraged an automated approach to highlight the effectiveness of drugs against the disease based on sentiment analysis of the text in the literature. Through this literature mining, we found dexamethasone, ivermectin, remdesivir, and others in the list of potential drugs for COVID-19 treatment. We highlighted hydroxychloroquine as an ineffective drug against COVID-19. We extracted disease-gene associations from the literature and, based on cosine similarity against the gold-standard DisGeNET data set, provided a confidence level for the associations between diseases and genes. We found 194 drug-PDB associations, which highlighted the large amount of work performed by the scientific community to understand the mechanism behind drug-target interactions and virus-host protein interaction mechanisms for coronavirus-related diseases. Surprisingly, we found few miRNAs related to COVID-19, indicating the primary focus of the scientific community is toward protein-based drugs rather than RNA-based drugs, though there have been successful RNA-based antiviral drugs. One such drug is Miravirsen, which binds miR-122 to prevent it from hybridizing with the RNA genome of HCV, depriving HCV of its essential cellular cofactor and blocking HCV replication [75]. We expect more research along these lines in the coming months.

Research Implications

Currently, we are facing the largest public health emergency since the 1918 influenza outbreak [76]. From the beginning of this outbreak, the scientific community has invested large amounts of effort to create vaccines and identify therapeutic solutions. Vaccines for SARS-CoV-2 might come too late to have any effect on the first wave of the COVID-19 pandemic [77]. However, vaccines might be useful in subsequent waves of COVID-19 or in a postpandemic scenario in which COVID-19 becomes a seasonal virus [77]. In this scenario, the

identification of drugs with good efficacy and minimal side effects is a rational goal that can be achieved in the near future to combat SARS-CoV-2 [48]. Although promising pharmacological results with repurposed drugs are emerging every day, unfortunately, no drug has been approved thus far for the treatment of COVID-19. Repurposed drugs are under investigation worldwide, many in preclinical and clinical stages [78]. With increasing information about SARS-CoV-2, along with publications about similar respiratory diseases (eg, pneumonia, SARS), it will be essential to investigate existing drugs that are already known to be effective against other respiratory diseases. As a prime example, dexamethasone, an FDA-approved drug, was known to be effective against pneumonia [79], respiratory failure [80], and other diseases. However, there was no evidence of its effectiveness against COVID-19 until its recent breakthrough in a clinical trial [55]. Although final approval of the drug is still pending, had it been investigated earlier, more lives could have been saved.

The research in this study is expected to support the scientific community and decision makers in identifying candidate drugs with proper evidence from the scientific literature. This will also help stakeholders explore existing drugs that are already known to be effective against other respiratory diseases. Although careful manual curation of the identified associations of biomedical entities is the ultimate goal, our novel approach estimates the effectiveness of drugs for coronavirus-related diseases based on natural language processing, sentiment analysis, and deep learning to help the scientific community shorten the potential list of drugs, ultimately saving time and resources.

Tool and Availability

We made our computational workflow and the resulting database an open source tool named COVID-19Base for use by the scientific community [81,82]. It not only identifies the terms and associations, but also highlights the relevant literature through its digital object identifier (DOI) so that any researcher using this tool can easily check the original source for more detailed information. As the number of scientific publications related to COVID-19 is constantly increasing, we will update the knowledge base on a monthly basis and integrate all recent updates in the knowledge base. COVID-19Base has already gone through its first transformation (from COVID-19Base 1.0 to COVID-19Base 2.0), as the CORD-19 data set was updated during the manuscript preparation phase. The earlier version of the CORD-19 data set contained about 44,000 papers, whereas the current version includes more than 138,000. The knowledge base materials and the source code of our computational approach are available on GitHub [83].

Limitations

Understandably, our findings as presented in the knowledge base may have some errors due to the inherent limitations of the methods and approaches adopted. This is why the identified inferences and associations are made available to users for review and a feedback mechanism is included in COVID-19Base.

Conclusions

We proposed a dictionary-based automated computational workflow to find the associations of six different thematic areas related to COVID-19/SARS-CoV-2 and other coronavirus-related diseases in humans. We prepared a knowledge base and made it available as a tool for the scientific community. We believe this knowledge base will help the research community explore the existing drugs and biomedical entities for coronavirus-related diseases, and the lessons learned before this outbreak will allow us to find an effective treatment for COVID-19.

Acknowledgments

The open access publication of this article was funded by the College of Science and Engineering, Hamad Bin Khalifa University.

Conflicts of Interest

None declared.

References

1. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos Solitons Fractals 2020 May;134:109761 [FREE Full text] [doi: 10.1016/j.chaos.2020.109761] [Medline: 32308258]
2. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 2020 Apr 24;368(6489):395-400 [FREE Full text] [doi: 10.1126/science.aba9757] [Medline: 32144116]
3. Heymann DL, Shindo N. COVID-19: what is next for public health? The Lancet 2020 Feb;395(10224):542-545. [doi: 10.1016/s0140-6736(20)30374-3]
4. de Wit E, Rasmussen AL, Falzarano D, Bushmaker T, Feldmann F, Brining DL, et al. Middle East respiratory syndrome coronavirus (MERS-CoV) causes transient lower respiratory tract infection in rhesus macaques. Proc Natl Acad Sci U S A 2013 Oct 08;110(41):16598-16603. [doi: 10.1073/pnas.1310744110] [Medline: 24062443]
5. Alam I, Kamau AA, Kulmanov M, Jaremko Ł, Arold ST, Pain A, et al. Functional Pangenome Analysis Shows Key Features of E Protein Are Preserved in SARS and SARS-CoV-2. Front Cell Infect Microbiol 2020;10:405 [FREE Full text] [doi: 10.3389/fcimb.2020.00405] [Medline: 32850499]

6. Muralidharan N, Sakthivel R, Velmurugan D, Gromiha MM. Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with SARS-CoV-2 protease against COVID-19. J Biomol Struct Dyn 2020 Apr 16:1-6. [doi: 10.1080/07391102.2020.1752802] [Medline: 32248766]
7. Stebbing J, Phelan A, Griffin I, Tucker C, Oechsle O, Smith D, et al. COVID-19: combining antiviral and anti-inflammatory treatments. The Lancet Infectious Diseases 2020 Apr;20(4):400-402. [doi: 10.1016/s1473-3099(20)30132-8]
8. Kandeel M, Al-Nazawi M. Virtual screening and repurposing of FDA approved drugs against COVID-19 main protease. Life Sci 2020 Jun 15;251:117627 [FREE Full text] [doi: 10.1016/j.lfs.2020.117627] [Medline: 32251634]
9. Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: a critical review. Curr Med Chem 2013 Jun 01;20(23):2839-2860. [doi: 10.2174/09298673113209990001] [Medline: 23651302]
10. Li Y, Wang C, Miao Z, Bi X, Wu D, Jin N, et al. ViRBase: a resource for virus-host ncRNA-associated interactions. Nucleic Acids Res 2015 Jan;43(Database issue):D578-D582 [FREE Full text] [doi: 10.1093/nar/gku903] [Medline: 25274736]
11. Tang D, Li B, Xu T, Hu R, Tan D, Song X, et al. VISDB: a manually curated database of viral integration sites in the human genome. Nucleic Acids Res 2020 Jan 08;48(D1):D633-D641 [FREE Full text] [doi: 10.1093/nar/gkz867] [Medline: 31598702]
12. Zhang Y, Zmasek C, Sun G, Larsen CN, Scheuermann RH. Hepatitis C Virus Database and Bioinformatics Analysis Tools in the Virus Pathogen Resource (ViPR). Methods Mol Biol 2019;1911:47-69. [doi: 10.1007/978-1-4939-8976-8_3] [Medline: 30593617]
13. Pickett B, Greer D, Zhang Y, Stewart L, Zhou L, Sun G, et al. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses 2012 Nov 19;4(11):3209-3226 [FREE Full text] [doi: 10.3390/v4113209] [Medline: 23202522]
14. Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset. ArXiv Preprint posted online on April 22, 2020. [Medline: 32510522]
15. Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res 2017 Jan 04;45(D1):D619-D625 [FREE Full text] [doi: 10.1093/nar/gkw1033] [Medline: 27799471]
16. Goodsell DS, Zardecki C, Di Costanzo L, Duarte JM, Hudson BP, Persikova I, et al. RCSB Protein Data Bank: Enabling biomedical research and drug discovery. Protein Sci 2020 Jan 29;29(1):52-65. [doi: 10.1002/pro.3730] [Medline: 31531901]
17. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res 2008 Jan 23;36(Database issue):D154-D158 [FREE Full text] [doi: 10.1093/nar/gkm952] [Medline: 17991681]
18. Schriml L, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 2019 Jan 08;47(D1):D955-D962 [FREE Full text] [doi: 10.1093/nar/gky1032] [Medline: 30407550]
19. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006 Jan 01;34(Database issue):D668-D672 [FREE Full text] [doi: 10.1093/nar/gkj067] [Medline: 16381955]
20. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016 Jan 04;44(D1):D1075-D1079 [FREE Full text] [doi: 10.1093/nar/gkv1075] [Medline: 26481350]
21. Aho AV, Corasick MJ. Efficient string matching. Commun ACM 1975 Jun;18(6):333-340. [doi: 10.1145/360825.360855]
22. Wang P, Hao T, Yan J, Jin L. Large-scale extraction of drug-disease pairs from the medical literature. Journal of the Association for Information Science and Technology 2017 Jun 06;68(11):2649-2661. [doi: 10.1002/asi.23876]
23. Zimek A, Schubert E, Kriegel H. A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analy Data Mining 2012 Aug 27;5(5):363-387. [doi: 10.1002/sam.11161]
24. Lloyd S. Least squares quantization in PCM. IEEE Trans Inform Theory 1982 Mar;28(2):129-137. [doi: 10.1109/tit.1982.1056489]
25. Han L. Using a dynamic K-means algorithm to detect anomaly activities. : IEEE; 2012 Presented at: Seventh International Conference on Computational Intelligence and Security; December 3-4, 2011; Hainan, China. [doi: 10.1109/cis.2011.233]
26. Lima M, Zarpelão BB, Sampaio LDH, Rodrigues JJPC, Abrão T, Proença ML. Anomaly detection using baseline and k-means clustering. : IEEE; 2010 Presented at: SoftCOM 2010, 18th International Conference on Software, Telecommunications and Computer Networks; 2010; Split, Dubrovnik, Croatia p. 305-309.
27. Lu W, Traore I. Unsupervised anomaly detection using an evolutionary extension of k-means algorithm. IJICS 2008;2(2):107. [doi: 10.1504/ijics.2008.018513]
28. Syarif I, Prugel-Bennett A, Wills G. Unsupervised Clustering Approach for Network Anomaly Detection. In: Benlamri R, editor. Networked Digital Technologies. Berlin, Heidelberg: Springer; 2012.
29. Yasami Y, Mozaffari SP. A novel unsupervised classification approach for network anomaly detection by k-Means clustering and ID3 decision tree learning methods. J Supercomput 2009 Oct 9;53(1):231-245. [doi: 10.1007/s11227-009-0338-x]
30. Le QV, Mikolov T. Distributed Representations of Sentences and Documents. ArXiv Preprint posted online on May 16, 2014. [FREE Full text]
31. Textblob: simplified text processing. URL: https://textblob.readthedocs.io/en/dev/ [accessed 2020-11-03]
32. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv Preprint posted online on January 16, 2013. [FREE Full text]

33. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 1972 Jan;28(1):11-21. [doi: 10.1108/eb026526]
34. Luhn HP. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. 1957 Oct;1(4):309-317. [doi: 10.1147/rd.14.0309]
35. Pasupa K, Sunhem W. A comparison between shallow and deep architecture classifiers on small dataset. 2016 Presented at: 8th International Conference on Information Technology and Electrical Engineering (ICITEE); 2016; Yogyakarta, Indonesia. [doi: 10.1109/iciteed.2016.7863293]
36. Feng S, Zhou H, Dong H. Using deep neural network with small dataset to predict material defects. 2019 Jan;162:300-310. [doi: 10.1016/j.matdes.2018.11.060]
37. Hahnloser RHR, Sarpeshkar R, Mahowald MA, Douglas RJ, Seung HS. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 2000 Jun 22;405(6789):947-951. [doi: 10.1038/35016072] [Medline: 10879535]
38. Karlik B, Olgac AV. Performance analysis of various activation functions in generalized MLP architectures of neural networks. International Journal of Artificial Intelligence and Expert Systems 2011;1(4):111-122 [FREE Full text]
39. Menon A, Mehrotra K, Mohan CK, Ranka S. Characterization of a Class of Sigmoid Functions with Applications to Neural Networks. Neural Networks 1996 Jul;9(5):819-835. [doi: 10.1016/0893-6080(95)00107-7]
40. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. 2010 Presented at: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics; 2010; Sardinia, Italy.
41. Kingma DP, Ba J. Adam: A method for stochastic optimization. ArXiv Preprint posted online on December 22, 2014. [FREE Full text]
42. Singhal A. Modern information retrieval: A brief overview. IEEE Data Eng Bull 2001;24(4):1-43.
43. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017 Jan 04;45(D1):D833-D839 [FREE Full text] [doi: 10.1093/nar/gkw943] [Medline: 27924018]
44. Wang J, Sheng W, Fang C, Chen Y, Wang J, Yu C, et al. Clinical manifestations, laboratory findings, and treatment outcomes of SARS patients. Emerg Infect Dis 2004 May;10(5):818-824 [FREE Full text] [doi: 10.3201/eid1005.030640] [Medline: 15200814]
45. Chen C, Chen C, Yan JT, Zhou N, Zhao JP, Wang DW. [Analysis of myocardial injury in patients with COVID-19 and association between concomitant cardiovascular diseases and severity of COVID-19]. Zhonghua Xin Xue Guan Bing Za Zhi 2020 Jul 24;48(7):567-571. [doi: 10.3760/cma.j.cn112148-20200225-00123] [Medline: 32141280]
46. Wang G, Wu C, Zhang Q, Wu F, Yu B, Lv J, et al. C-Reactive Protein Level May Predict the Risk of COVID-19 Aggravation. Open Forum Infect Dis 2020 May;7(5):ofaa153 [FREE Full text] [doi: 10.1093/ofid/ofaa153] [Medline: 32455147]
47. Akgun E, Tuzuner MB, Sahin B, Kilercik M, Kulah C, Cakiroglu HN, et al. Altered molecular pathways observed in naso-oropharyngeal samples of SARS-CoV-2 patients. medRxiv Preprint posted online on May 18, 2020. [FREE Full text] [doi: 10.1101/2020.05.14.20102558]
48. Potì F, Pozzoli C, Adami M, Poli E, Costa LG. Treatments for COVID-19: emerging drugs against the coronavirus. Acta Biomed 2020 May 11;91(2):118-136. [doi: 10.23750/abm.v91i2.9639] [Medline: 32420936]
49. Walls AC, Park Y, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 2020 Apr 16;181(2):281-292.e6 [FREE Full text] [doi: 10.1016/j.cell.2020.02.058] [Medline: 32155444]
50. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 2020 Apr 16;181(2):271-280.e8 [FREE Full text] [doi: 10.1016/j.cell.2020.02.052] [Medline: 32142651]
51. Zang R, Gomez Castro MF, McCune BT, Zeng Q, Rothlauf PW, Sonnek NM, et al. TMPRSS2 and TMPRSS4 promote SARS-CoV-2 infection of human small intestinal enterocytes. Sci Immunol 2020 May 13;5(47) [FREE Full text] [doi: 10.1126/sciimmunol.abc3582] [Medline: 32404436]
52. Wicik Z, Eyileten C, Jakubik D, Simões SN, Martins Jr DC, Pavão R, et al. ACE2 interaction networks in COVID-19: a physiological framework for prediction of outcome in patients with cardiovascular risk factors. BioRxiv Preprint posted online on October 9, 2020. [doi: 10.1101/2020.05.13.094714]
53. He R, Lu Z, Zhang L, Fan T, Xiong R, Shen X, et al. The clinical course and its correlated immune status in COVID-19 pneumonia. J Clin Virol 2020 Jun;127:104361 [FREE Full text] [doi: 10.1016/j.jcv.2020.104361] [Medline: 32344320]
54. D'Amico F, Baumgart DC, Danese S, Peyrin-Biroulet L. Diarrhea During COVID-19 Infection: Pathogenesis, Epidemiology, Prevention, and Management. Clin Gastroenterol Hepatol 2020 Jul;18(8):1663-1672 [FREE Full text] [doi: 10.1016/j.cgh.2020.04.001] [Medline: 32278065]
55. RECOVERY Collaborative Group, Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, et al. Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report. N Engl J Med 2020 Jul 17:7273 [FREE Full text] [doi: 10.1056/NEJMoa2021436] [Medline: 32678530]
56. Chaccour C, Hammann F, Ramón-García S, Rabinovich NR. Ivermectin and COVID-19: Keeping Rigor in Times of Urgency. Am J Trop Med Hyg 2020 Jun;102(6):1156-1157 [FREE Full text] [doi: 10.4269/ajtmh.20-0271] [Medline: 32314704]

57. Caly L, Druce JD, Catton MG, Jans DA, Wagstaff KM. The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro. Antiviral Res 2020 Jun;178:104787 [FREE Full text] [doi: 10.1016/j.antiviral.2020.104787] [Medline: 32251768]
58. Sujan M. Use of Ivermectin: Hope held out, caution called for. The Daily Star. 2020. URL: https://www.thedailystar.net/frontpage/news/use-ivermectin-hope-held-out-caution-called-1914041 [accessed 2020-06-26]
59. Al-Tawfiq JA, Al-Homoud AH, Memish ZA. Remdesivir as a possible therapeutic option for the COVID-19. Travel Med Infect Dis 2020 Mar;34:101615 [FREE Full text] [doi: 10.1016/j.tmaid.2020.101615] [Medline: 32145386]
60. Tchesnokov E, Feng J, Porter D, Götte M. Mechanism of Inhibition of Ebola Virus RNA-Dependent RNA Polymerase by Remdesivir. Viruses 2019 Apr 04;11(4):326 [FREE Full text] [doi: 10.3390/v11040326] [Medline: 30987343]
61. Warren TK, Jordan R, Lo MK, Ray AS, Mackman RL, Soloveva V, et al. Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. Nature 2016 Mar 17;531(7594):381-385 [FREE Full text] [doi: 10.1038/nature17180] [Medline: 26934220]
62. NIH Clinical Trial Shows Remdesivir Accelerates Recovery from Advanced COVID-19. 2020. URL: https://www.niaid.nih.gov/news-events/nih-clinical-trial-shows-remdesivir-accelerates-recovery-advanced-covid-19 [accessed 2020-06-26]
63. Clinical trials related to COVID-19. URL: https://clinicaltrials.gov/ct2/results?cond=COVID-19 [accessed 2020-06-26]
64. Hagar M, Ahmed HA, Aljohani G, Alhaddad OA. Investigation of Some Antiviral -Heterocycles as COVID 19 Drug: Molecular Docking and DFT Calculations. Int J Mol Sci 2020 May 30;21(11):3922 [FREE Full text] [doi: 10.3390/ijms21113922] [Medline: 32486229]
65. Grein J, Ohmagari N, Shin D, Diaz G, Asperges E, Castagna A, et al. Compassionate Use of Remdesivir for Patients with Severe Covid-19. N Engl J Med 2020 Jun 11;382(24):2327-2336 [FREE Full text] [doi: 10.1056/NEJMoa2007016] [Medline: 32275812]
66. Wang Y, Zhang D, Du G, Du R, Zhao J, Jin Y, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. Lancet 2020 May 16;395(10236):1569-1578 [FREE Full text] [doi: 10.1016/S0140-6736(20)31022-9] [Medline: 32423584]
67. FDA cautions against use of hydroxychloroquine or chloroquine for COVID-19 outside of the hospital setting or a clinical trial due to risk of heart rhythm problems. URL: https://www.fda.gov/drugs/drug-safety-and-availability/fda-cautions-against-use-hydroxychloroquine-or-chloroquine-covid-19-outside-hospital-setting-or [accessed 2020-05-12]
68. Cannon CP, Braunwald E, McCabe CH, Rader DJ, Rouleau JL, Belder R, et al. Intensive versus Moderate Lipid Lowering with Statins after Acute Coronary Syndromes. N Engl J Med 2004 Apr 08;350(15):1495-1504. [doi: 10.1056/nejmoa040583]
69. Dashti-Khavidaki S, Khalili H. Considerations for Statin Therapy in Patients with COVID-19. Pharmacotherapy 2020 May 04;40(5):484-486 [FREE Full text] [doi: 10.1002/phar.2397] [Medline: 32267560]
70. Castiglione V, Chiriacò M, Emdin M, Taddei S, Vergaro G. Statin therapy in COVID-19 infection. Eur Heart J Cardiovasc Pharmacother 2020 Jul 01;6(4):258-259 [FREE Full text] [doi: 10.1093/ehjcvp/pvaa042] [Medline: 32347925]
71. Coronavirus Response - Active Support for Hospitalised Covid-19 Patients (CRASH-19). 2020. URL: https://clinicaltrials.gov/ct2/show/NCT04343001 [accessed 2020-06-26]
72. Atorvastatin as Adjunctive Therapy in COVID-19 (STATCO19). 2020. URL: https://clinicaltrials.gov/ct2/show/NCT04380402 [accessed 2020-06-26]
73. Deliwala S, Abdulhamid S, Abusalih MF, Al-Qasmi MM, Bachuwa G. Encephalopathy as the Sentinel Sign of a Cortical Stroke in a Patient Infected With Coronavirus Disease-19 (COVID-19). Cureus 2020 May 14;12(5):e8121 [FREE Full text] [doi: 10.7759/cureus.8121] [Medline: 32426200]
74. Alexander S, Armstrong JF, Davenport AP, Davies JA, Faccenda E, Harding SD, et al. A rational roadmap for SARS-CoV-2/COVID-19 pharmacotherapeutic research and development: IUPHAR Review 29. Br J Pharmacol 2020 Nov;177(21):4942-4966 [FREE Full text] [doi: 10.1111/bph.15094] [Medline: 32358833]
75. Ottosen S, Parsley TB, Yang L, Zeh K, van Doorn L, van der Veer E, et al. Antiviral Activity and Preclinical and Clinical Resistance Profile of Miravirsen, a Novel Anti-Hepatitis C Virus Therapeutic Targeting the Human Factor miR-122. Antimicrob Agents Chemother 2014 Nov 10;59(1):599-608. [doi: 10.1128/aac.04220-14]
76. Shi Y, Wang Y, Shao C, Huang J, Gan J, Huang X, et al. COVID-19 infection: the perspectives on immune responses. Cell Death Differ 2020 May 23;27(5):1451-1454 [FREE Full text] [doi: 10.1038/s41418-020-0530-3] [Medline: 32205856]
77. Amanat F, Krammer F. SARS-CoV-2 Vaccines: Status Report. Immunity 2020 Apr 14;52(4):583-589 [FREE Full text] [doi: 10.1016/j.immuni.2020.03.007] [Medline: 32259480]
78. Rosa S, Santos WC. Clinical trials on drug repositioning for COVID-19 treatment. Rev Panam Salud Publica 2020;44:e40 [FREE Full text] [doi: 10.26633/RPSP.2020.40] [Medline: 32256547]
79. Remmelts HH, Meijvis SC, Heijligenberg R, Rijkers GT, Oosterheert JJ, Bos WJW, et al. Biomarkers define the clinical response to dexamethasone in community-acquired pneumonia. J Infect 2012 Jul;65(1):25-31. [doi: 10.1016/j.jinf.2012.03.008] [Medline: 22410382]
80. da Costa DE, Nair AK, Pai MG, Al Khusaiby SM. Steroids in full term infants with respiratory failure and pulmonary hypertension due to meconium aspiration syndrome. Eur J Pediatr 2001 Mar 14;160(3):150-153. [doi: 10.1007/s004310000678] [Medline: 11277374]
81. COVID-19Base 2.0. URL: https://covid-19base.hbku.edu.qa/search [accessed 2020-11-05]

82. COVID-19Base 2.0. URL: http://covid-19base.buet.ac.bd/search [accessed 2020-11-05]
83. COVID-19Base GitHub. URL: https://github.com/JunaedYounusKhan51/COVID-19Base [accessed 2020-11-03]

Abbreviations

ACE2: angiotensin-converting enzyme 2
CORD-19: COVID-19 Open Research Dataset
CRP: C-reactive protein
DNN: deep neural network
DO: Disease Ontology
FDA: Food and Drug Administration
HCV: hepatitis C virus
HGNC: HUGO Gene Nomenclature Committee
miRNA: micro ribonucleic acid
ncRNA: noncoding ribonucleic acid
NIAID: National Institute of Allergy and Infectious Diseases
PDB: Protein Data Bank
POS: part-of-speech
ReLU: rectified linear unit
tf–idf: term frequency–inverse document frequency

Edited by G Eysenbach, C Lovis; submitted 26.06.20; peer-reviewed by A Civit, A Sarafi Nejad; comments to author 21.08.20; revised version received 23.08.20; accepted 06.09.20; published 10.11.20.

Please cite as:
Khan JY, Khondaker MTI, Hoque IT, Al-Absi HRH, Rahman MS, Guler R, Alam T, Rahman MS
Toward Preparing a Knowledge Base to Explore Potential Drugs and Biomedical Entities Related to COVID-19: Automated Computational Approach
JMIR Med Inform 2020;8(11):e21648
URL: http://medinform.jmir.org/2020/11/e21648/
doi:10.2196/21648
PMID:33055059

Original Paper

Prediction of COVID-19 Severity Using Chest Computed Tomography and Laboratory Measurements: Evaluation Using a Machine Learning Approach

Daowei Li[1*], MD; Qiang Zhang[2*], MD; Yue Tan[3], MD; Xinghuo Feng[4], MD; Yuanyi Yue[3], MD; Yuhan Bai[5], MD; Jimeng Li[6], MD; Jiahang Li[6], MD; Youjun Xu[7], MD; Shiyu Chen[8], MD; Si-Yu Xiao[9], PhD; Muyan Sun[9], PhD; Xiaona Li[10], MD; Fang Zhu[11], MD

[1]Department of Radiology, The People's Hospital of China Medical University & The People's Hospital of Liaoning Province, Shenyang, China

[2]Department of Pulmonary and Critical Care Medicine, Shengjing Hospital of China Medical University, Shenyang, China

[3]Department of Gastroenterology, Shengjing Hospital of China Medical University, Shenyang, China

[4]Department of Intensive Care Unit, The People's Hospital of Yicheng City, Yicheng, China

[5]The First Clinical Department, China Medical University, Shenyang, China

[6]The Second Clinical Department, China Medical University, Shenyang, China

[7]Department of Radiology, The People's Hospital of Yicheng City, Yicheng, China

[8]Department of Laboratory Medicine, The People's Hospital of Yicheng City, Yicheng, China

[9]Intanx Life (Shanghai) Co, Ltd, Shanghai, China

[10]School of Fundamental Sciences, China Medical University, Shenyang, China

[11]Department of Cardiovascular Ultrasound, The People's Hospital of China Medical University & The People's Hospital of Liaoning Province, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Fang Zhu, MD
Department of Cardiovascular Ultrasound
The People's Hospital of China Medical University & The People's Hospital of Liaoning Province
No 33, Wenyi Road
Shenhe District
Shenyang, 110016
China
Phone: 86 2483283333
Email: zfmoon024@163.com

Abstract

Background: Most of the mortality resulting from COVID-19 has been associated with severe disease. Effective treatment of severe cases remains a challenge due to the lack of early detection of the infection.

Objective: This study aimed to develop an effective prediction model for COVID-19 severity by combining radiological outcome with clinical biochemical indexes.

Methods: A total of 46 patients with COVID-19 (10 severe, 36 nonsevere) were examined. To build the prediction model, a set of 27 severe and 151 nonsevere clinical laboratory records and computerized tomography (CT) records were collected from these patients. We managed to extract specific features from the patients' CT images by using a recently published convolutional neural network. We also trained a machine learning model combining these features with clinical laboratory results.

Results: We present a prediction model combining patients' radiological outcomes with their clinical biochemical indexes to identify severe COVID-19 cases. The prediction model yielded a cross-validated area under the receiver operating characteristic (AUROC) score of 0.93 and an F_1 score of 0.89, which showed a 6% and 15% improvement, respectively, compared to the models based on laboratory test features only. In addition, we developed a statistical model for forecasting COVID-19 severity based on the results of patients' laboratory tests performed before they were classified as severe cases; this model yielded an AUROC score of 0.81.

Conclusions: To our knowledge, this is the first report predicting the clinical progression of COVID-19, as well as forecasting severity, based on a combined analysis using laboratory tests and CT images.

(JMIR Med Inform 2020;8(11):e21604) doi:10.2196/21604

KEYWORDS

COVID-19; severe case prediction; computerized tomography; machine learning; CT; scan; detection; prediction; model

Introduction

In December 2019, an epidemic of pneumonia caused by a newly identified coronavirus (SARS-CoV-2) emerged in China and has been spreading worldwide ever since [1]. According to the World Health Organization, to date, the COVID-19 pandemic has affected more than 200 countries worldwide, causing global panic and contributing to fears of market recession and mass unemployment. The novel virus causing COVID-19 was identified to have originated from the Orthocoronavirinae subfamily, the same subfamily as SARS-CoV and MERS-CoV [2], and it was thus officially named SARS-CoV-2. This virus might invade the human airway epithelial cells by binding to the angiotensin-converting enzyme 2 receptor (ACE2), in a mechanism similar to that of SARS-CoV [3,4].

The clinical features of COVID-19 are atypical, ranging from mild systematic symptoms, including intermittent fever (83%) and lower respiratory tract reactions such as cough (61%), to less common ones such as shortness of breath (14.5%), muscle ache (18.6%), headache (11.8%), and diarrhea (6.1%) [1,5]. Some patients with COVID-19 might develop severe complications such as acute renal failure (2.1%), acute respiratory distress syndrome (ARDS, 8.9%), or shock (2.2%), and some might even die (3.7%) [1,6]. The clinical and epidemiological spectrum of COVID-19 is quite diverse and is still not fully understood. Previous reports have suggested that the whole world's population is generally prone to COVID-19 [7]. Nevertheless, older patients who have underlying diseases such as cerebral infarction, chronic obstructive pulmonary disease, bronchiectasis, or diabetes are more prone to severe pneumonia, respiratory failure, septic shock, or even death caused by multiple organ failure [6].

SARS-CoV-2 is highly infectious and can be primarily transmitted through direct or indirect contact, droplets, and aerosol. Diagnosis of COVID-19 usually involves a combination of the patient's travel history, clinical symptoms, and radiological and biochemical findings. Patchy ipsilateral pulmonary consolidations are visible on a computerized tomography (CT) scan initially, during the early course of COVID-19. As the infection progresses, the consolidations are reduced and appear as bilateral ground-glass opacities, marking the prominent radiological features of COVID-19 [8]. The "white lung" radiograph, a characteristic finding suggesting that the patient urgently requires oxygen inhalation, has only been observed in a few critical patients with ARDS [9-11]. Other biochemical index changes associated with a COVID-19 diagnosis include lymphopenia, increased C-reactive protein and lactate dehydrogenase (LDH) levels, and thrombocytopenia [5].

Antiviral medication and glucocorticoids are most commonly used for the clinical treatment of COVID-19, with antibacterial medication sought when bacterial co-infection is detected [12]. Given the insufficient clinical trial data for the safety and efficacy of remdesivir and chloroquine, there is still no persuasive evidence for effective medicine for the treatment of COVID-19 [13]. It is noteworthy that approximately 11.5% of all reported patients with COVID-19 developed severe illness characterized as ARDS. These patients were transferred to an intensive care unit, as they required mechanical ventilation and even extracorporeal membrane pulmonary oxygenation (ECMO), the efficacy of which is very limited according to a retrospective study, wherein 5 of 6 patients receiving ECMO eventually died [14,15]. In fact, the mortality rate of severely ill patients with a confirmed diagnosis of COVID-19 is 60%, indicating the importance of early detection and prediction of COVID-19 severity [14,15]. However, at present, it is a critical challenge to identify a patient with COVID-19 who might require intensive care before certain clinical symptoms are observed. Therefore, there is an urgent need to develop an effective prediction or forecasting model for patients with COVID-19.

Our study aimed to address this challenge: we developed a prediction model for COVID-19 clinical progression, by combining radiological outcome based on CT scans with biochemical indexes. To extract essential features from CT scans, we segmented the lungs from the CT volumetric images by using a deep convolutional neural network (CNN). Finally, we also developed a model to forecast COVID-19 severity based on the results of the patients' laboratory tests before the patients were classified as severe cases. To our knowledge, this is the first study to report a prediction model for assessing COVID-19 severity by combining radiological outcomes with clinical biochemical indexes. We believe that our prediction model will shed light on predicting disease severity for all patients with COVID-19.

Methods

Patient Information

We collected samples from 46 patients who visited People's Hospital of Yicheng City between January 16, 2020, and March 4, 2020, and were diagnosed with COVID-19 according to the Chinese Government Diagnosis and Treatment Guideline (Trial 5th version; Medicine, 2020). For a confirmed diagnosis of COVID-19, nucleic acid was extracted from sputum or throat swab samples using a nucleic acid extractor (EX3600, Shanghai Zhijiang Biotechnology Co.) and a nucleic acid extraction reagent (No. P20200201, Shanghai Zhijiang Biotechnology Co.).

Fluorescence-based quantitative polymerase chain reaction (PCR; ABI7500) and SARS-CoV-2 nucleic acid detection kit (triple fluorescence PCR, No. P20200203, Shanghai Zhijiang Biotechnology Co.) were used for nucleic acid detection. This kit uses a one-step reverse transcription–PCR combined with Taqman technology to detect RNA-dependent RNA polymerase (*RdRp*), envelope (E), and nucleocapsid (N) genes. PCR results were concluded to be positive if (1) *RdRp* gene was positive (cycle threshold [Ct]<43) and either E or N gene was positive (Ct<43), or (2) if two sequential tests for *RdRp* were positive and those for E and N genes were negative. The 46 study patients with COVID-19 were classified into 2 types: (1) nonsevere, comprising patients showing mild symptoms without radiological manifestations of pneumonia, fever, or respiratory tract symptoms with radiological manifestations of pneumonia, and (2) severe, comprising patients meeting any of the following criteria—respiratory rate ≥30 breaths/min, pulse oxygen saturation ≤93% in resting state, partial pressure of arterial oxygen ≤300 mm Hg (1 mm Hg=0.133 kPa), respiratory failure requiring mechanical ventilation, shock incidence, and admission to intensive care unit with other organ failure. In total, 10 patients were categorized as severe cases and 36, as nonsevere cases. The last follow-up of these patients was on March 10, 2020.

Ethics Approval

Approval for studies on CT screening and clinical test results was obtained from the Medical Ethics Committee of The People's Hospital of Yicheng City, China (2020Yc002)

Data Collection

We collected and reviewed clinical information of 46 patients with COVID-19 after admission, including clinical signs and symptoms, comorbidities, travel history, laboratory tests, and CT scans. To consolidate all patients' records into a single table, missing records for a given day were noted as "NA" (not available). In all, we obtained 178 records (27 severe and 151 nonsevere cases) from 105 different laboratory tests and chest CT images. Note that throughout the clinical course, each patient had more than one record variably classified as severe or nonsevere. Patients with at least one severe record were classified as severe cases.

Data Processing and Statistical Analysis

We identified 44 laboratory tests that had more than 50% missing values (NA), and we then imputed the NAs with the mean values. Related laboratory tests were identified based on the criterion that the *P* value (Mann-Whitney U test) between the severe and nonsevere groups is smaller than .05. In all, we found that 36 laboratory tests were related to the detection of COVID-19 severity. The patients' CT images were processed using a pretrained CNN with a U-Net structure [16] to segment the lung lobes from the background. The intensities were then normalized to grayscale for all patients before further analysis. We then analyzed the intensities of the 3D CT volumes within lung masks to obtain CT features for each record.

Severity Prediction Models

Prediction models were developed to predict patient severity based on laboratory and CT signatures collected at corresponding dates. Each patient record was considered a sample for a model; as a result, 178 samples were evaluated using those models. Before using model prediction, we used random forest importance score, mutual information, and fold change as possible approaches to select important model features while avoiding potential overfitting. We found mutual information to be the most robust approach. We considered different candidate machine learning models, including random forest classifier, gradient boost classifier, XGB classifier, logistic classifier, and supported vector machine. Random forest was found to be the best classifier, and model parameters were optimized using a genetic algorithm (Tree-Based Pipeline Optimization Tool). The area under the curve of the receiver operating characteristic (AUROC) and F_1 scores were used to evaluate model accuracy considering the dataset imbalance. All models were trained with 5-fold cross-validation with stratified train-test splits that preserve the percentage of samples in severe and nonsevere groups. All cross-validated results were averaged over 20 runs.

Severity Forecasting Models

Forecasting models were built to forecast patient severity based on laboratory and CT signatures collected from nonsevere cases at admission. In these models, instead of the patients' records, the patients themselves were considered as samples to build forecasting relationships. CT records were not collected as frequently as laboratory tests were performed, and initial, nonsevere CT records were not available for 3 severe cases. Therefore, we built two separate random forest models based on CT features and laboratory tests with 7 and 10 severe cases, respectively. Other model details were identical to those of the severity prediction models.

Results

Overview of Study Patients

We collected clinical data of 46 patients with COVID-19 who were admitted at the People's Hospital of Yicheng City, between mid-January and early-March 2020. We recorded 305 biochemical test results from 105 different tests, based on the clinical reports of all 46 study patients (Multimedia Appendix 1). General patient information is shown in Table 1. The general trend that older patients with COVID-19 tend to develop more systemic symptoms was not observed in our study [17]. However, patients with comorbidities, especially diabetes and hypertension, tended to develop more severe symptoms than others. Moreover, patients with severe COVID-19 typically experienced fatigue, anorexia, malaise, chest congestion, and shortness of breath.

Table 1. Characteristics and symptoms of study patients.

Characteristic	Values		
	All cases (N=46)	Severe cases (n=10)	Nonsevere cases (n=36)
Age in years, mean (range)	48.8 (24-71)	56.8 (33-71)	46.5 (24-71)
Sex, n (%)			
Male	25 (54)	6 (60)	19 (53)
Female	21 (46)	4 (40)	17 (47)
Exposure, n (%)			
Wuhan	24 (52)	5 (50)	19 (53)
Family	4 (9)	2 (20)	2 (6)
Community	5 (11)	0 (0)	5 (14)
None	13 (28)	3 (30)	10 (28)
Comorbidity, n (%)			
Hypertension	11 (24)	5 (50)	6 (17)
Cardiovascular disease	6 (13)	2 (20)	4 (11)
Chronic liver disease	3 (7)	2 (20)	1 (3)
Diabetes	5 (11)	3 (30)	2 (6)
Leukoderma	1 (2)	0 (0)	1 (3)
Chronic kidney disease	1 (2)	0 (0)	1 (3)
Hyperuricemia	1 (2)	0 (0)	1 (3)
Chronic lung disease	2 (4)	0 (0)	2 (6)
Symptoms , n (%)			
Dry Cough	28 (61)	6 (60)	22 (61)
Cough with phlegm	9 (20)	2 (20)	7 (19)
Fever			
High	8 (17)	3 (30)	5 (14)
Mid	20 (43)	4 (40)	10 (28)
Mild	14 (30)	3 (30)	17 (47)
Fatigue	25 (54)	9 (90)	16 (44)
Anorexia	33 (72)	9 (90)	24 (67)
Malaise	34 (74)	10 (100)	24 (67)
Headache	7 (15)	3 (30)	4 (11)
Nausea	1 (2)	0 (0)	1 (3)
Diarrhea	5 (11)	2 (20)	3 (8)
Dyspnea	1 (2)	1 (10)	0 (0)
Chest congestion	16 (35)	5 (50)	11 (31)
Shortness of breath after activity	19 (41)	6 (60)	13 (36)

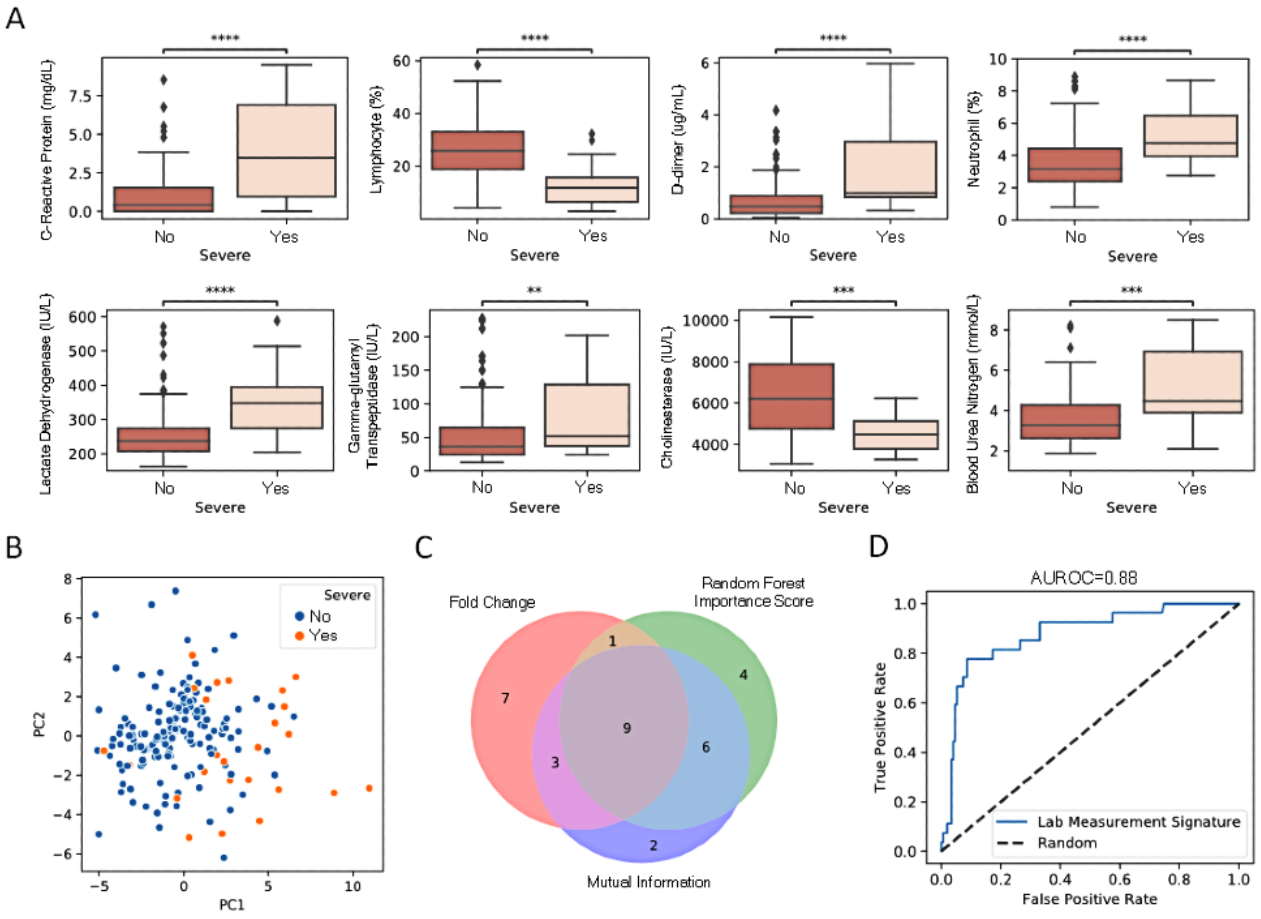
In all, 52% (24/46) patients had a travel history to or from Wuhan within the past 1 month, and 20% (9/46) patients had clear exposure history in the local city (Table 1). According to the patients' medical records, 80% of the severe cases had one or more comorbidities, whereas only 34% of the nonsevere cases had comorbidities. This finding is consistent with that of previous studies [18]. Moreover, 81% (37/46) patients had cough and only 20% (9/46) patients reported sputum production. Fever was the most common symptom; however, severe cases (7/10, 70%) had a higher proportion of mid- to high-grade fever (ie, >38.9°C) than the nonsevere cases (15/36, 42%). More than half of the patients (25/46, 54%) experienced fatigue, and approximately three-quarters of them had anorexia (33/46, 72%) or malaise (34/46, 74%); these symptoms were observed in almost all severe cases (fatigue, 9/10, 90%; anorexia, 9/10, 90%; and malaise, 10/10, 100%). Headache, nausea, diarrhea, and

dyspnea were rarely observed in both severe and nonsevere cases. Moreover, less than half of all patients reported chest congestion (16/46, 35%) or shortness of breath after activity (19/46, 41%), and these symptoms were approximately 20% more common in severe cases than in nonsevere cases.

Prediction Based on Laboratory Tests

Data processing yielded 61 laboratory tests results, 36 of which were significantly related to severity. Eight related laboratory tests that showed the largest fold change are illustrated in Figure 1A. Among these tests, D-dimer, LDH, and lymphocytes were found to be associated with mortality risk [17]. Principal component analysis results clearly showed separation between the severe and nonsevere groups, indicating that the COVID-19–related laboratory tests can be used to identify disease severity (Figure 1B). To build a statistical model to predict severity, we first selected the most important laboratory tests to avoid overfitting. Three different approaches—fold change, random forest importance score, and mutual information—were considered the top-ranking laboratory features. In fact, the three approaches led to very similar ranking, and the top features obtained from mutual information resulted in the largest intersection with those obtained from the other two approaches (Figure 1C). This finding suggests that mutual information is the most robust feature selector among the three abovementioned approaches; therefore, we used mutual information to select laboratory features to be used in the model. We used a random forest model with hyperparameters optimized by a genetic algorithm (see Methods) to predict severity based on laboratory features. Our results suggested that the prediction accuracy does not further increase with an increase in the number of laboratory features beyond 12. As a result, a signature of the top 12 laboratory features was considered, and the corresponding model yielded a cross-validated AUROC score of 0.88 and an F_1 score of 0.69 (Figure 1D).

Figure 1. Correlation of laboratory tests with COVID-19 severity. (A) Top-8 laboratory tests ranked by fold change. (B) Principal component (PC) analysis of all laboratory tests. (C) Venn diagram of the top features selected by 3 different approaches: random forest importance score, mutual information, and fold change. (D) Area under receiver operating characteristic of classification using a signature of 12 laboratory tests. The asterisk annotations denote the following: * 1.00e-02<P≤5.00e-02, ** 1.00e-03<P≤1.00e-02, *** 1.00e-04<P≤1.00e-03, **** for P≤1.00e-04.



Extraction of CT Features

To extract CT features, we first segmented the lungs from the CT volumetric images using a deep CNN, U-Net. Because the CNN was pretrained with several annotated datasets, including a COVID-19 dataset from MedSeg [16], we directly transferred the trained CNN to segment CT images of the study patients. The CT slices acquired across a clinical course of a patient are shown in Figure 2A. Right after onset, the patient was diagnosed with nonsevere disease, with no apparent opacity visible in lung CT. The patient was classified as severe on Day 4, and this continued for 2 weeks thereafter, with increasing amounts of ground-glass opacity and patchy consolidation. The ground-glass opacity and consolidation started to fade away from Day 27, and on Day 30, the patient was confirmed to be asymptomatic. As seen in Figure 2A, the opacity of the segmented lung lobes

is associated with disease severity. The opacity can be represented by the intensity distribution within the segmented lung volumes (Figure 2B). Note that only the slices in the middle are shown in Figure 2A for illustration purposes; all slices were considered, however, to determine intensity distribution. As the symptoms became severe, the background became increasingly opaque, as indicated by the peak locations and peak heights of the intensity distribution. The distribution also changed from unimodal to bimodal. Therefore, we considered peak location and height as well as the first four moments of the intensity distribution (ie, mean, standard deviation, skewness, and kurtosis) as CT features. Since the intensity distribution can become bimodal, we also added the Otsu threshold to reflect the bimodality and entropy to supplement standard deviation. Three exemplary CT features observed along the clinical course of the patient are shown in Figure 2D. Thus, Otsu threshold is an excellent predictor for severity based on visualization. We then analyzed all 178 CT records and determined the corresponding intensity distributions (Figure 3A and 3B). We found that distributions of severe and nonsevere cases were in direct contrast in terms of peak height and skewness. Principal component analysis also showed improved separation between the 2 groups (Figure 3C). Among the 8 CT features examined, peak location and entropy were not significantly related to severity, whereas all the other 6 CT features showed a statistically significant relation (Figure 3C). Mean and standard deviation, as well as skewness and kurtosis, were highly correlated; therefore, standard deviation and kurtosis were not considered as CT features.

Figure 2. Computed tomography (CT) feature extraction. (A) Segmented lung images from the middle CT slice for a patient with a full course of COVID-19 from nonsevere to severe and then from severe to nonsevere. The patient's severe records are presented in red color. (B) Intensity histograms of the volume CT within segmented lung masks for five consecutive records of the patient. (C) Peak location and Otsu threshold features from the intensity histogram on Day 18. (D) Variation of 3 different CT features along the course of the disease.

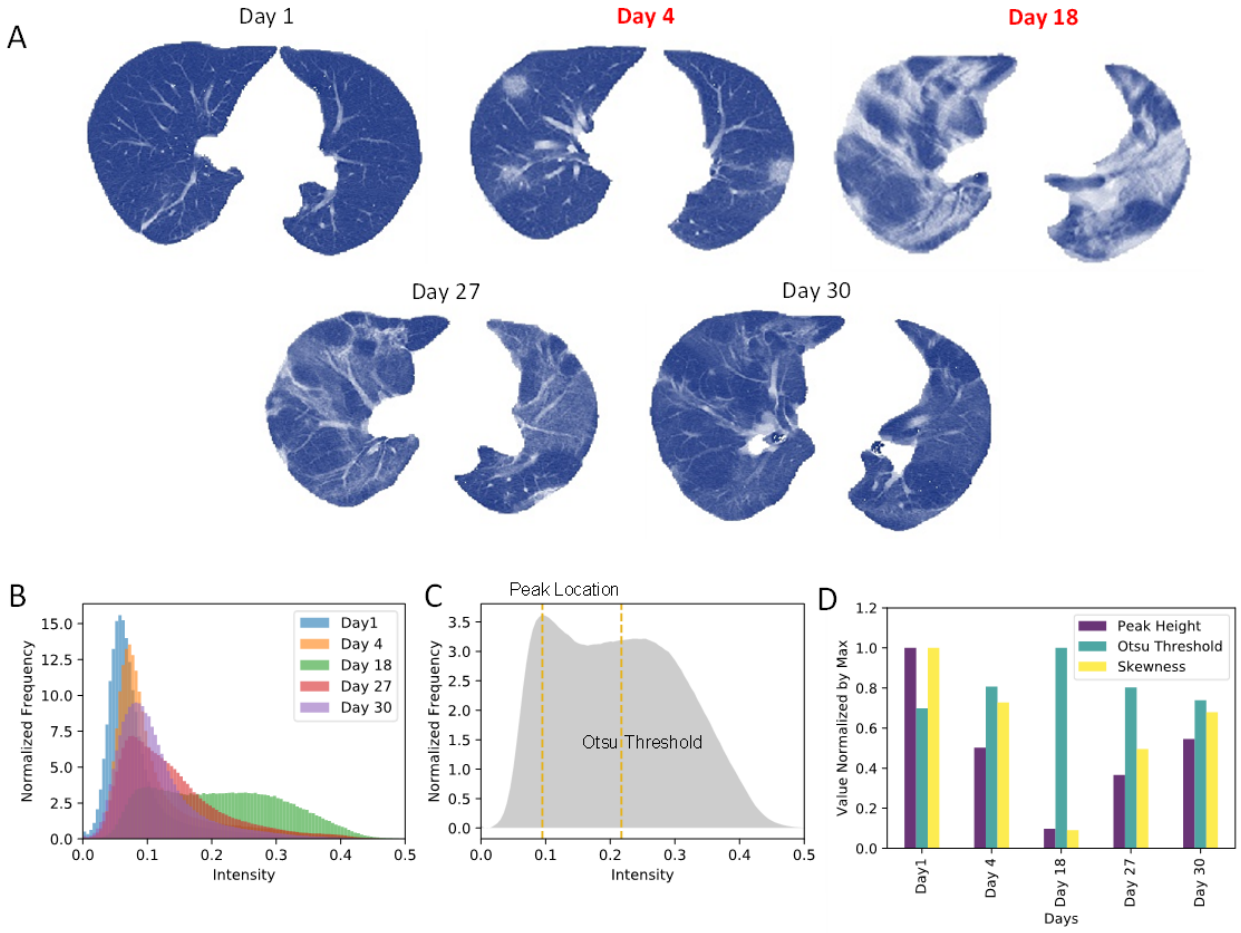
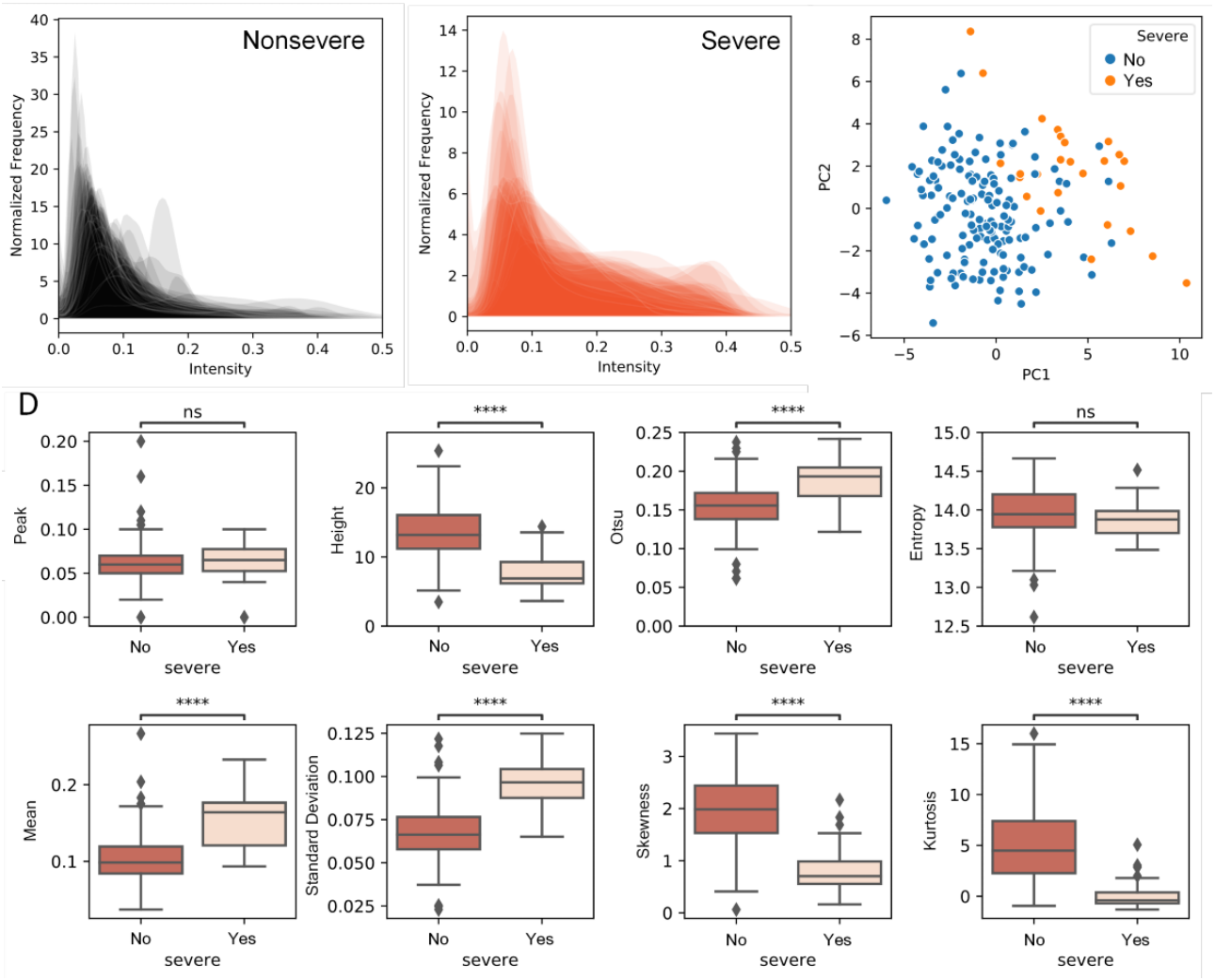


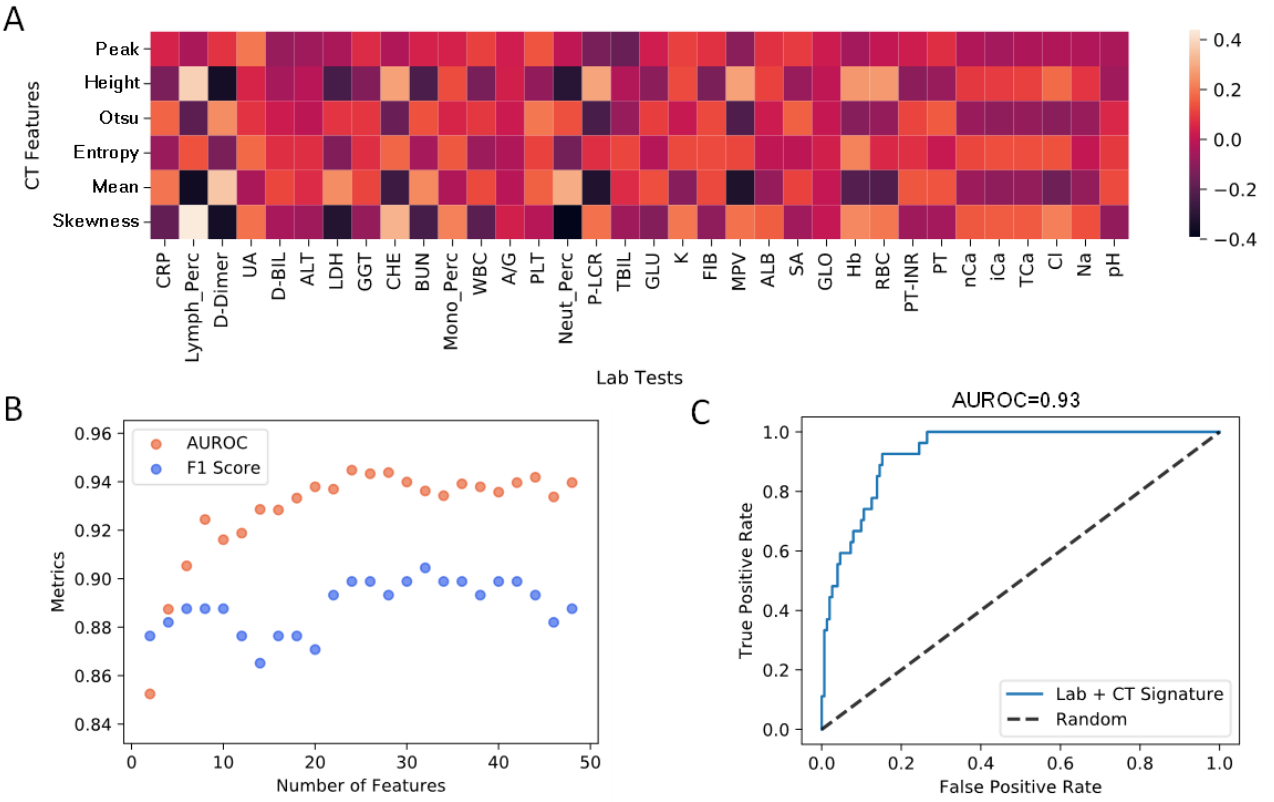
Figure 3. Computed tomography (CT) intensity distribution and extracted features of patients with COVID-19. (A) Intensity distribution of CT volumes from nonsevere cases. (B) Intensity distribution of CT volumes from severe cases. (C) Principal component analysis of all CT features. (D) All CT features between severe and nonsevere groups. "Peak" stands for peak location, and "height" stands for peak height. The asterisk annotations denote the following: * 1.00e-02<P≤5.00e-02, ** 1.00e-03<P≤1.00e-02, *** 1.00e-04<P≤1.00e-03, **** P≤1.00e-04.



Prediction Based on CT and Laboratory Features

The CT feature extraction enables quantitative prediction with signatures of both CT and laboratory features. We first analyzed the Spearman correlation between the CT and laboratory features (Figure 4A). Most features were not significantly correlated; however, lymphocyte, neutrophil, D-dimer, and platelet–large cell ratio showed good correlation with several CT features. Similarly, we used mutual information to select features to be used in the model. We used a random forest model with optimized hyperparameters to predict severity from CT and laboratory features. We selected a signature of 16 features from the feature number analysis (Figure 4B). The corresponding prediction model yielded a cross-validated AUROC score of 0.93 and an F_1 score of 0.81 (Figure 4C), which are considerably improved from the corresponding scores of the model with laboratory tests only (Figure 1D). The signature includes CT peak height, CT intensity mean, CT intensity skewness, CT Otsu threshold, lymphocyte percentage, gamma-glutamyl transpeptidase, LDH, C-reactive protein, white blood cell, D-dimer, cholinesterase, neutrophil percentage, hemoglobin, tricyclic antidepressant, albumin, and chloride.

Figure 4. Prediction based on computed tomography (CT) and laboratory features. (A) Spearman correlation heatmap between CT and laboratory features. “Peak” stands for peak location, and “height” stands for peak height. A summary table describing all CT and laboratory features and their abbreviations is provided in Multimedia Appendix 2. (B) Model accuracy metrics with an increased number of features. (C) Area under receiver operating characteristic of classification using a signature of 15 CT and laboratory features.



Forecasting Disease Severity

Forecasting disease severity has significant clinical importance, as it allows clinicians to better prepare for treatment course. In addition to predicting severity based on CT and laboratory signatures, we also developed a statistical model to forecast severity from patient records upon admission when they were considered nonsevere. Although CT features are excellent predictors of severity, they are not as good for forecasting, yielding an AUROC of 0.68. In contrast, the random forest model based on laboratory tests yielded an AUROC of 0.81, indicating excellent forecasting predictability (Figure 5A). Other metrics considered for forecasting are presented in Table 2. This statistical model comprised 8 laboratory tests, among which lymphocyte and neutrophil counts (percentage) showed the highest fold change. In addition to comorbidity, we identified 8 laboratory tests that could be used for severity forecasting: individual counts of lymphocyte, neutrophil, monocyte, and eosinophil; red blood cell distribution width; hemoglobin; procalcitonin; and platelet–large cell ratio.

Figure 5. COVID-19 severity forecasted using the prediction model. (A) Forecasting severity using patient's nonsevere records noted upon admission. (B) Laboratory tests showing a significant relation to the severity forecast.

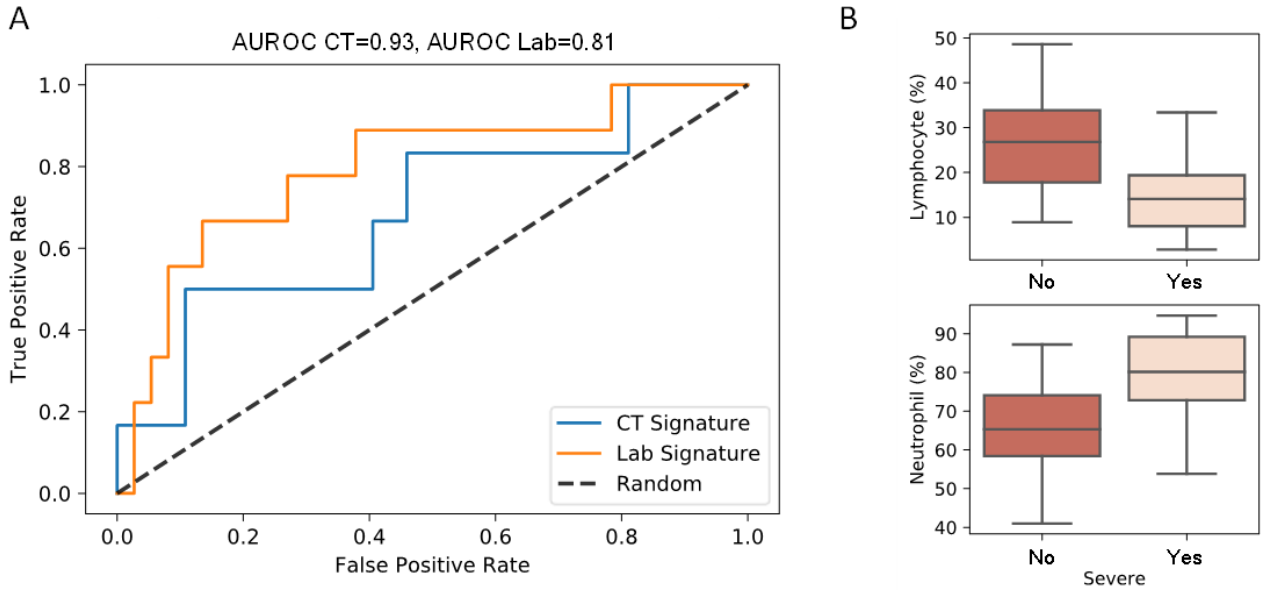


Table 2. Metrics of prediction and forecasting models. Mean and standard deviation values across 5 cross-validation splits are shown. AUROC: area under the receiver operating characteristics.

Features	Prediction model		Forecasting model	
	Laboratory only, mean (SD)	Laboratory and CT[a], mean (SD)	CT only, mean (SD)	Laboratory only, mean (SD)
Precision	0.75 (0.2)	0.82 (0.05)	0.55 (0.22)	0.61 (0.23)
Recall	0.7 (0.15)	0.79 (0.1)	0.56 (0.23)	0.61 (0.11)
AUROC[b] Score	0.86 (0.1)	0.93 (0.03)	0.68 (0.22)	0.81 (0.14)
F_1 Score	0.69 (0.17)	0.81 (0.05)	0.56 (0.22)	0.60 (0.16)
Accuracy	0.87 (0.04)	0.88 (0.03)	0.78 (0.12)	0.83 (0.06)

[a]CT: computed tomography.

[b]AUROC: area under the receiver operating characteristic.

Discussion

In this study, we collected clinical records from 46 patients with COVID-19 (27 severe and 151 nonsevere records) and developed a prediction model using a combination of radiological outcomes and clinical biochemical indexes, to identify disease severity. Using the model thus developed, we successfully achieved an AUROC score of 0.93 to identify the patient's severity status. Furthermore, we established a model for forecasting disease severity based on the combined features recorded before the patients were classified as severe cases, resulting in an AUROC score of 0.81.

In the history of confrontation between human beings and pathogens, humans have always been prone to losing the battle when the development of effective medicine or vaccine is extremely difficult owing to the high variability of the pathogenic genome, such as in the case of influenza virus, HIV, or SARS. Even though the reported mortality rate of COVID-19 (1.4% [5]) is not as high as that of SARS (10% [19]), individuals with underlying health conditions such as hypertension, cardiovascular disease, chronic kidney disease, and diabetes (2.89-, 3.84-, 2.22-, and 2.65-fold higher risk, respectively [20]) are much more vulnerable to COVID-19. Approximately half of the patients with COVID-19 are above 50 years of age [5]; these patients are much more likely to develop severe symptoms such as those characterized by ARDS or multiple organ failure. Moreover, the significant need for early prediction of clinical progression has aroused much attention worldwide, yet it remains to be fully addressed.

Many studies highlight the potential hallmarks of COVID-19. Biochemical and radiological outcomes are the most widely recognized indexes in clinical treatment and decision making [21]; these include interleukin-6 level [22], lymphocyte count, neutrophil-to-lymphocyte ratio [23], aspartate aminotransferase level [24], and ground glass opacity on CT scan images [11,13,25]. An artificial intelligence tool focuses on early detection by screening publicly available radiological results of patients with COVID-19 with an accuracy of 86.7% [26-28]. Another recent study developed a system based on deep learning models to quantify the infectious areas in the lungs of patients with COVID-19, to predict the severity of clinical course [29]. A prognostic model based on the XGBoost algorithm with a reported accuracy greater than 90% used 3 biochemical features, including LDH, to predict the mortality rate and clinical

outcomes [30], whereas another machine learning framework based on random forest, decision tree, and support vector machine used 3 different clinical features, including aminotransferase, for early prediction of clinical severity [31]. However, the accuracy of the latter model was 70%-80% when an adequate dataset was not available, as only incomplete information from 53 patients was used for the analysis. Interestingly, all published research for the prediction of clinical severity focused on either biochemical or radiological indexes only. To our knowledge, our study presents the first prognostic model using both biochemical indexes and CT scan results based on neural network and deep learning, which significantly improves the predictive capability as suggested by an AUROC score of 0.93. The limitations of this study include a limited sample size and incomplete information about the patients' past medical history—challenges often encountered by clinicians in critical and urgent scenarios. Our future work will be focused on increasing sample size and improving data quality.

Conclusions

In conclusion, the course of clinical progression might be clearer with the application of our model, and we believe our effort could provide useful opinions for early identification of severely ill patients. Thus, advanced interventions could be applied to potentially reduce mortality rates and alleviate the health care burden regarding the management of COVID-19 cases.

Acknowledgments

We thank all patients and donors involved in this study. We appreciate the assistance received from Intanx Life Co. Ltd. (Shanghai) in data processing and consulting.

Authors' Contributions

XL and FZ designed the research study and collected patient samples; DL, QZ, YT, Y, YB, Jimeng Li, Jiahang Li, YX, SX, and MS performed the research; DL, QZ, YT, SX, and MS analyzed the data; DL, QZ, YT, XL, and FZ wrote the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical laboratory data for study patients.
[XLSX File (Microsoft Excel File), 168 KB - medinform_v8i11e21604_app1.xlsx]

Multimedia Appendix 2
Summary of all clinical laboratory features as described in Figure 4A.
[XLSX File (Microsoft Excel File), 13 KB - medinform_v8i11e21604_app2.xlsx]

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506. [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, China Novel Coronavirus Investigating Research Team. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020 Feb 20;382(8):727-733 [FREE Full text] [doi: 10.1056/NEJMoa2001017] [Medline: 31978945]
3. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020 Feb 22;395(10224):565-574 [FREE Full text] [doi: 10.1016/S0140-6736(20)30251-8] [Medline: 32007145]
4. Wu F, Zhao S, Yu B, Chen Y, Wang W, Song Z, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020 Mar;579(7798):265-269 [FREE Full text] [doi: 10.1038/s41586-020-2008-3] [Medline: 32015508]
5. Fang Z, Yi F, Wu K, Lai K, Sun X, Zhong N. Clinical characteristics of coronavirus disease 2019 (COVID-19): an updated systematic review. medRxiv. Preprint posted online on March 12, 2020. [doi: 10.1101/2020.03.07.20032573]
6. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020 Feb 15;395(10223):507-513 [FREE Full text] [doi: 10.1016/S0140-6736(20)30211-7] [Medline: 32007143]
7. Liu K, Chen Y, Lin R, Han K. Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. J Infect 2020 Jun;80(6):e14-e18 [FREE Full text] [doi: 10.1016/j.jinf.2020.03.005] [Medline: 32171866]
8. Carotti M, Salaffi F, Sarzi-Puttini P, Agostini A, Borgheresi A, Minorati D, et al. Chest CT features of coronavirus disease 2019 (COVID-19) pneumonia: key points for radiologists. Radiol Med 2020 Jul;125(7):636-646. [doi: 10.1007/s11547-020-01237-4] [Medline: 32500509]
9. Pan Y, Guan H. Imaging changes in patients with 2019-nCov. Eur Radiol 2020 Jul;30(7):3612-3613. [doi: 10.1007/s00330-020-06713-z] [Medline: 32025790]

10. Pan Y, Guan H, Zhou S, Wang Y, Li Q, Zhu T, et al. Initial CT findings and temporal changes in patients with the novel coronavirus pneumonia (2019-nCoV): a study of 63 patients in Wuhan, China. Eur Radiol 2020 Jun;30(6):3306-3309 [FREE Full text] [doi: 10.1007/s00330-020-06731-x] [Medline: 32055945]
11. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis 2020 Apr;20(4):425-434 [FREE Full text] [doi: 10.1016/S1473-3099(20)30086-4] [Medline: 32105637]
12. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA 2020 Mar 17;323(11):1061-1069 [FREE Full text] [doi: 10.1001/jama.2020.1585] [Medline: 32031570]
13. Liu F, Zhang Q, Huang C, Shi C, Wang L, Shi N, et al. CT quantification of pneumonia lesions in early days predicts progression to severe illness in a cohort of COVID-19 patients. Theranostics 2020;10(12):5613-5622 [FREE Full text] [doi: 10.7150/thno.45985] [Medline: 32373235]
14. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med 2020 May;8(5):475-481. [doi: 10.1016/S2213-2600(20)30079-5] [Medline: 32105632]
15. Henry BM. COVID-19, ECMO, and lymphopenia: a word of caution. Lancet Respir Med 2020 Apr;8(4):e24 [FREE Full text] [doi: 10.1016/S2213-2600(20)30119-3] [Medline: 32178774]
16. Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp 2020 Aug 20;4(1):50 [FREE Full text] [doi: 10.1186/s41747-020-00173-2] [Medline: 32814998]
17. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet 2020 Mar 28;395(10229):1054-1062 [FREE Full text] [doi: 10.1016/S0140-6736(20)30566-3] [Medline: 32171076]
18. Chen R, Liang W, Jiang M, Guan W, Zhan C, Wang T, Medical Treatment Expert Group for COVID-19. Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China. Chest 2020 Jul;158(1):97-105 [FREE Full text] [doi: 10.1016/j.chest.2020.04.010] [Medline: 32304772]
19. Parry J. WHO warns that death rate from SARS could reach 10%. BMJ 2003 May 10;326(7397):999 [FREE Full text] [doi: 10.1136/bmj.326.7397.999/a] [Medline: 12742900]
20. Wang X, Fang X, Cai Z, Wu X, Gao X, Min J, et al. Comorbid chronic diseases and acute organ injuries are strongly correlated with disease severity and mortality among COVID-19 patients: a systemic review and meta-analysis. Research (Wash D C) 2020;2020:2402961 [FREE Full text] [doi: 10.34133/2020/2402961] [Medline: 32377638]
21. Li J, Chen Z, Nie Y, Ma Y, Guo Q, Dai X. Identification of symptoms prognostic of COVID-19 severity: multivariate data analysis of a case series in Henan province. J Med Internet Res. Preprint posted online June 16, 2020. [doi: 10.2196/preprints.19636]
22. Liu F, Li L, Xu M, Wu J, Luo D, Zhu Y, et al. Prognostic value of interleukin-6, C-reactive protein, and procalcitonin in patients with COVID-19. J Clin Virol 2020 Jun;127:104370 [FREE Full text] [doi: 10.1016/j.jcv.2020.104370] [Medline: 32344321]
23. Ma Y, Shi N, Fan Y, Wang J, Zhao C, Li G, et al. Predictive value of the neutrophil-to-lymphocyte ratio(NLR) for diagnosis and worse clinical course of the COVID-19: findings from ten provinces in China. SSRN Journal. Preprint posted online April 14, 2020. [doi: 10.2139/ssrn.3569838]
24. Ji D, Zhang D, Chen Z, Xu Z, Zhao P, Zhang M, et al. Clinical characteristics predicting progression of COVID-19. SSRN Journal. Preprint posted online Feb 20, 2020. [doi: 10.2139/ssrn.3539674]
25. Yu Q, Wang Y, Huang S, Liu S, Zhou Z, Zhang S, et al. Multicenter cohort study demonstrates more consolidation in upper lungs on initial CT increases the risk of adverse clinical outcome in COVID-19 patients. Theranostics 2020;10(12):5641-5648 [FREE Full text] [doi: 10.7150/thno.46465] [Medline: 32373237]
26. Pereira RM, Bertolini D, Teixeira LO, Silla CN, Costa YM. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Comput Methods Programs Biomed 2020 Oct;194:105532 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105532] [Medline: 32446037]
27. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020 Jun;121:103792 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103792] [Medline: 32568675]
28. Ko H, Chung H, Kang WS, Kim KW, Shin Y, Kang SJ, et al. COVID-19 Pneumonia Diagnosis Using a Simple 2D Deep Learning Framework With a Single Chest CT Image: Model Development and Validation. J Med Internet Res 2020 Jun 29;22(6):e19569 [FREE Full text] [doi: 10.2196/19569] [Medline: 32568730]
29. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M. Lung infection quantification of COVID-19 in CT images with deep learning. arXiv. Preprint posted online Mar 10, 2020.
30. Yan L, Zhang H, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2020 May;2(5):283-288. [doi: 10.1038/s42256-020-0180-7]

31. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials and Continua 2020 Mar;63(1):537-551. [doi: 10.32604/cmc.2020.010691]

Abbreviations

ARDS: acute respiratory distress syndrome
AUROC: area under the receiver operating characteristic
CNN: convolutional neural network
CT: computerized tomography
Ct: cycle threshold
ECMO: extracorporeal membrane pulmonary oxygenation
LDH: lactate dehydrogenase
PCR: polymerase chain reaction
RT: reverse transcription

Edited by G Eysenbach; submitted 18.06.20; peer-reviewed by D Du, J Liu; comments to author 01.08.20; revised version received 10.08.20; accepted 21.09.20; published 17.11.20.

Please cite as:
Li D, Zhang Q, Tan Y, Feng X, Yue Y, Bai Y, Li J, Li J, Xu Y, Chen S, Xiao SY, Sun M, Li X, Zhu F
Prediction of COVID-19 Severity Using Chest Computed Tomography and Laboratory Measurements: Evaluation Using a Machine Learning Approach
JMIR Med Inform 2020;8(11):e21604
URL: http://medinform.jmir.org/2020/11/e21604/
doi:10.2196/21604
PMID:33038076

Original Paper

Applying eHealth for Pandemic Management in Saudi Arabia in the Context of COVID-19: Survey Study and Framework Proposal

Abdullah Alsharif[1], PhD

Department of Management Information Systems, College of Business Administration-Yanbu, Taibah University, Medinah, Saudi Arabia

Corresponding Author:
Abdullah Alsharif, PhD
Department of Management Information Systems
College of Business Administration-Yanbu
Taibah University
14 Khlaf Alkhfigee Street 8314
Iskan
Medinah, PO Box 344
Saudi Arabia
Phone: 966 542578585
Email: alsharifa@taibahu.edu.sa

Abstract

Background: The increased frequency of epidemics such as Middle East respiratory syndrome, severe acute respiratory syndrome, Ebola virus, and Zika virus has created stress on health care management and operations as well as on relevant stakeholders. In addition, the recent COVID-19 outbreak has been creating challenges for various countries and their respective health care organizations in managing and controlling the pandemic. One of the most important observations during the recent outbreak is the lack of effective eHealth frameworks for managing and controlling pandemics.

Objective: The aims of this study are to review the current National eHealth Strategy of Saudi Arabia and to propose an integrated eHealth framework that can be effective for managing health care operations and services during pandemics.

Methods: A questionnaire-based survey was administered to 316 health care professionals to review the current national eHealth framework of Saudi Arabia and identify the objectives, factors, and components that are key for managing and controlling pandemics. Purposive sampling was used to collect responses from diverse experts, including physicians, technical experts, nurses, administrative experts, and pharmacists. The survey was administered at five hospitals in Saudi Arabia by forwarding the survey link using a web-based portal. A sample population of 350 was achieved, which was filtered to exclude incomplete and ineligible samples, giving a sample of 316 participants.

Results: Of the 316 participants, 187 (59.2%) found the current eHealth framework to be ineffective, and more than 50% of the total participants stated that the framework lacked some essential components and objectives. Additional components and objectives focusing on using eHealth for managing information, creating awareness, increasing accessibility and reachability, promoting self-management and self-collaboration, promoting electronic services, and extensive stakeholder engagement were considered to be the most important factors by more than 80% of the total participants.

Conclusions: Managing pandemics requires an effective and efficient eHealth framework that can be used to manage various health care services by integrating different eHealth components and collaborating with all stakeholders.

(JMIR Med Inform 2020;8(11):e19524) doi:10.2196/19524

KEYWORDS

COVID-19; eHealth framework; infectious disease; pandemic; eHealth; public health

Introduction

Background

eHealth can be basically defined as the application of information and communications technology (ICT) in the health care sector [1]. eHealth is considered to be one of the major developments of the past few decades and has revolutionized the operation of health care services. Various studies [2-4] have indicated the positive impact of eHealth approaches in improving health care service delivery, minimizing operational costs, increasing process efficiency, and most importantly,

managing health care information. Focusing on the concept of "application" derived from the definition in [1], the success of eHealth depends on how it is applied and managed in different areas of health care operation. To improve the process of eHealth implementation, different frameworks were developed by different countries (eg, Austria in the European Union) and health care organizations according to their needs and specifications [5,6]. However, these frameworks and documents mainly focus on examples and present outlines (high level designs); meanwhile, they offer very little information to guide the process of development [1]. A few studies [7-9] have attempted to provide frameworks for strategic planning and implementation of eHealth strategies. However, these studies are mostly related to normal health care operations. In a few studies [10,11], frameworks were developed for health care management during pandemics; however, these studies are limited in scope, focusing on frameworks for assessing preparedness and readiness but not for managing health care operations during pandemics. These studies focused on readiness strategies, such as resource management, finances, and vaccination; none of them highlighted an operational framework for pandemics that includes eHealth approaches or stakeholders' roles and responsibilities. In addition, these studies were distributed across different regions and may not be applicable in all cases.

National eHealth frameworks have been formulated based on the vision and selected objectives targeting health care management and dissemination of general health care services [6,12]. These frameworks may not be effective in dealing with pandemics, which may require sudden changes in health care policies, strategies, and information management. For instance, various myths and pieces of misinformation about COVID-19 were observed to be circulating over the internet and proved to be challenging for governments and health care officials to address [13]. In addition, pandemics have been identified more frequently in recent years, including severe acute respiratory syndrome (SARS), H7N9 influenza, Zika virus, Ebola virus, Middle East respiratory syndrome (MERS), and COVID-19. Considering these unexpected risks and other health care challenges, sudden changes in health care policies and strategies may be inevitable. In these conditions, creating awareness among the public, disseminating legitimate information, and making changes to health care services such as access and delivery are aspects of eHealth that can be useful in managing pandemics. eHealth can be one of the most effective operations during pandemics, as it can enable remote management of health care operations and services. Focusing on the gaps identified in relation to the management of health care operations during pandemics, and considering these challenges and the possibilities of using eHealth approaches to pandemic management, the aims of this paper are to review the existing national eHealth framework of Saudi Arabia and to propose an integrated eHealth framework for managing pandemics in Saudi Arabia.

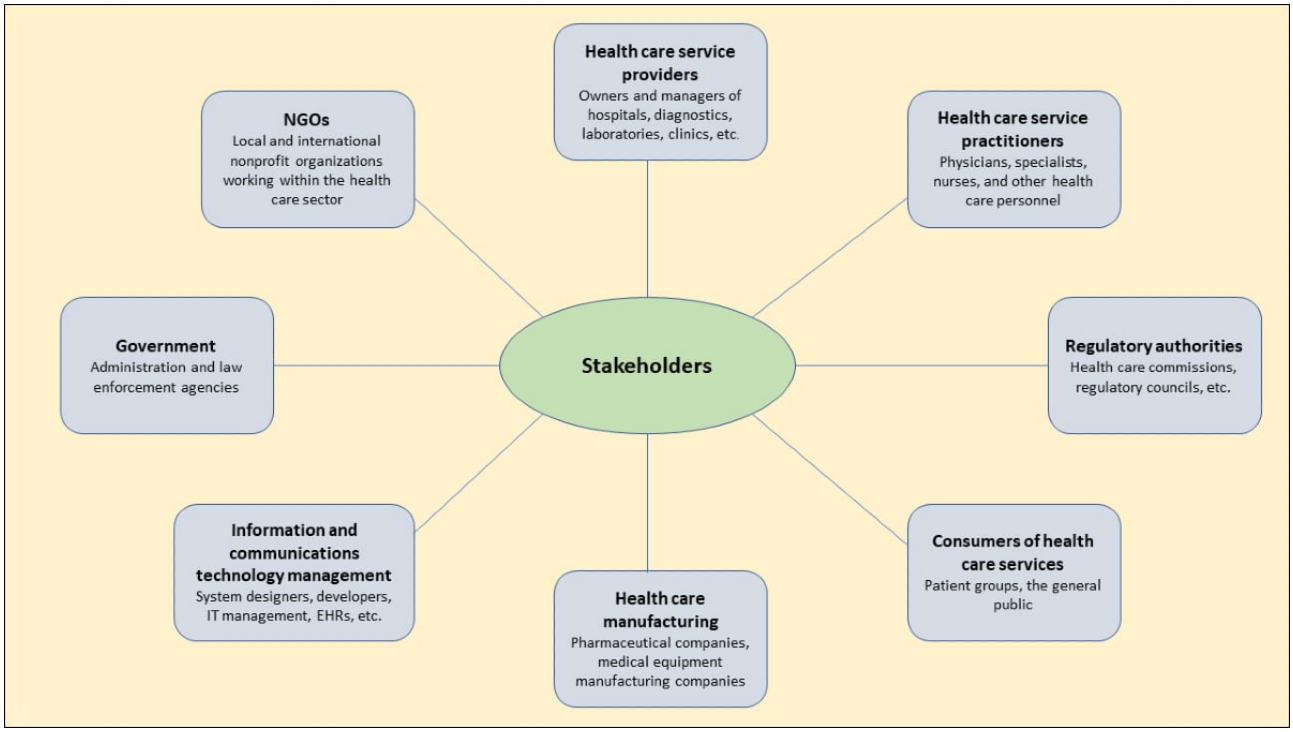
Literature Review

According to the World Health Organization (WHO) [14], of its 194 member countries, only 58% have an eHealth strategy, and only 55% have established legislation in relation to eHealth data. The frameworks developed by these countries were based on their needs and considered various contexts, such as finance, resources, and technological capabilities. Because few countries have an eHealth framework and proper eHealth legislation, it can be challenging for countries to coordinate and manage health care operations during pandemics, when there is a need for collaboration among countries to manage the pandemic. However, it is equally important to manage health care operations internally and control the rapid spread of disease and infection during pandemics. Different health care components may need to be effectively managed during pandemics. Accordingly, this section provides a review of the literature regarding various components of eHealth frameworks and their applicability in pandemics according to the guidelines prescribed by the WHO [15].

Stakeholders

Stakeholders are considered to be among the important components of pandemic management processes. Usually, a pandemic requires greater participation of various stakeholders in health care management, as shown in Figure 1; this is applicable not only to health care practitioners but also to society as a whole [16]. Therefore, the stakeholders in pandemic times include society, where every individual shares the responsibility for containing the spread of infectious diseases. However, there is a need to define the roles and responsibilities of all stakeholders, including health care workers, research organizations, health care equipment manufacturers, and pharmaceutical companies, to prepare for pandemics [17]. Health care service providers and practitioners have the responsibilities of planning and delivering health care services, managing health care resources, etc. Regulatory authorities may need to monitor the health care service operations across various hospitals and formulate standards and regulations to be adopted by health care service providers and practitioners. Accordingly, manufacturing organizations should focus on meeting the growing needs for medicines, equipment, and other resources, while the information and communications management unit should focus on managing web-based health care operations and creating public awareness. Governments should monitor all health care operations and administer these at different levels, including cities, towns, and villages, in both public and private hospitals to prevent contamination and risk of infection. Pandemics may have a serious impact on people's mental health. Anxiety, depression, and stress are illnesses that can result from fear about pandemics or under preventive measures such as isolation and social distancing [18,19]. Therefore, it is essential to provide support and awareness to the people. eHealth can be effective in this context, as support, services, and information can be remotely shared among people using various applications. Stakeholders usually rely on less timely and traditional sources of disease surveillance; therefore, there is a need for timely and reliable pandemic intelligence using effective communication technologies [20].

Figure 1. Stakeholders in health care management. EHR: electronic health record; IT, information technology; NGO: nongovernmental organization.



Operations Management

Operations management is another important area to be considered during pandemics. Available resources such as finance, medical supplies, equipment, devices, nurses, and practitioners need to be effectively managed, resulting in maximum efficiency of health care operations [21]. Large upfront investments in diagnosis, tests, and treatment were identified to result in the most efficient use of resources during pandemics [22]. In addition, pharmaceutical company operations related to research and development must be managed for maximum output. eHealth can play an important role in all these health-related operational activities by effectively disseminating quality information across the supply chain. In addition, operations such as diagnosis, disease surveillance, monitoring, tracing, and tracking can be effectively managed through mobile apps, which can limit the number of patients visiting hospitals and prevent the spread of infection.

Information Management

Information management is one of the most important aspects to be considered during pandemics. The successful management of a pandemic mostly relies on how information is managed and shared among stakeholders, which can help them take timely actions. The increasing number of myths related to COVID-19 is an example of poor information management, and rapid spread of incorrect information through social media platforms can result in serious damage, such as destruction of 5G towers in the United Kingdom [15] and drinking raw alcohol to prevent transmission of SARS-CoV-2 in Iran [23]. Effective monitoring and tracking and, most importantly, creating awareness about novel diseases and the precautions to be taken by the responsible authorities are key to the operational management and containment of infectious disease.

Strategy Development

The next set of factors focuses on the approach for developing strategies. These factors include strategic context; experiences from previous events and research; formulating a mission and objectives; identifying the target components; analyzing opportunities and gaps; developing strategies; implementation; monitoring and reviewing the approaches; and making necessary changes to approaches after deployment. Strategic context is related to the main area the approach is targeting. It may be related to containing infectious disease, reducing the spread of misinformation, or any other specific context related to pandemics [24]. Experiences from the past, such as approaches adopted by different governments to control pandemics such as SARS, MERS, Zika virus, and Ebola virus, can be used to identify the context. The mission statement is the foundation for developing the approach to managing pandemics. It must clearly define the purpose of the approach and instill a sense of motivation and hope among all the stakeholders [25] involved in containing pandemics. It should inspire and present a message of contributing to the delivery of the best health care services to people who are affected by pandemics through integrated clinical practice, education, research, and effective resource management to contain the pandemic.

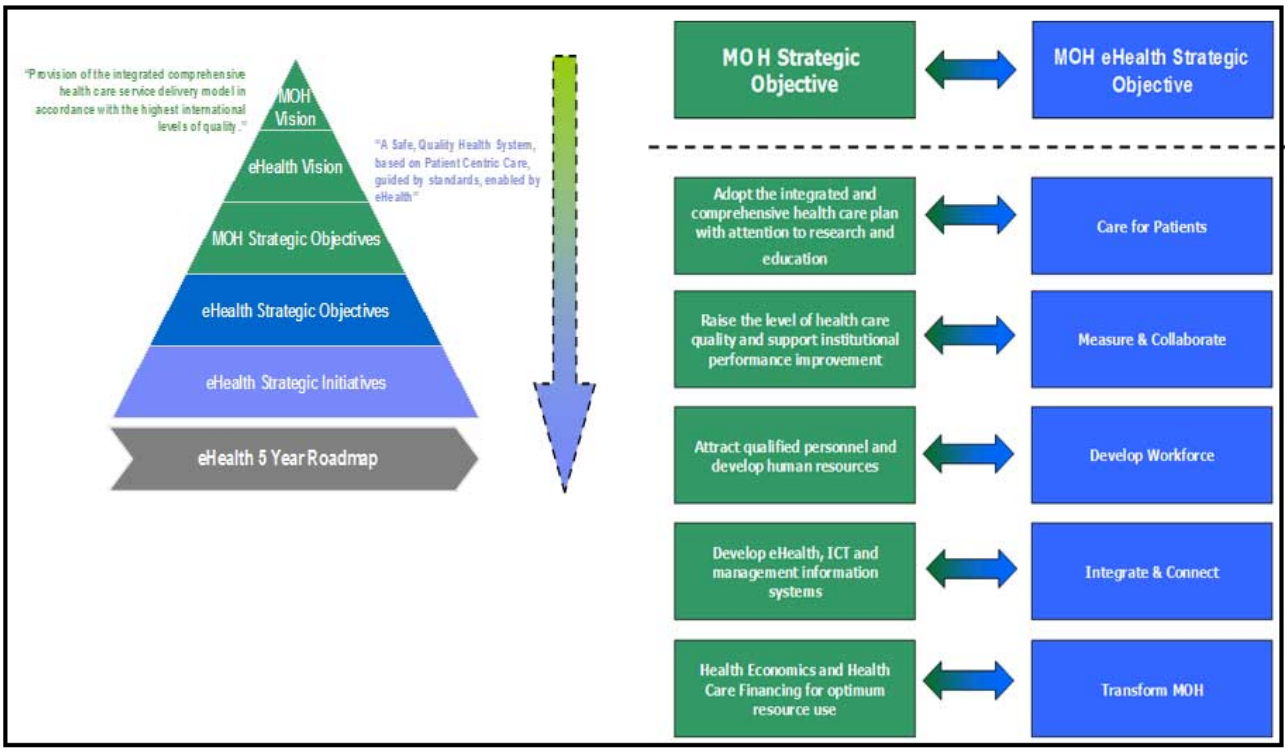
Formulation of Objectives

The next step focuses on formulating the objectives. The objectives define the set of goals or activities that need to be achieved. These can include statements related to developing the workforce, developing research activities, improving health care services and delivery, etc. [25]. Next, the components that need to be targeted are identified; these can include different areas and activities related to health care management [24]. Based on the objectives, opportunities and gaps for achieving the specified objectives must be identified. Based on the identified components, objectives, and available resources, the

necessary strategies for managing pandemics must be developed and implemented. The overall approach being adopted for managing pandemics must be monitored and evaluated. Based on the evaluation, the approach can be updated by making necessary changes in areas that are not effective. These guidelines can be used by different countries to develop eHealth strategies, as the basic components that must be considered are outlined. However, the success of the framework depends on how the strategies are formulated and implemented, which mainly involves the application of ICT to the health care framework.

As part of this study, the National eHealth Strategy of Saudi Arabia (shown in Figure 2) is reviewed in the context of its applicability to managing pandemics. The framework provides the details and objectives of the approach, which are usually updated every five years. However, the current strategy is unclear and reflective in relation to various components, and it lacks a clear approach and process, supporting the findings of [1]. Necessary components such as information management, operations management, and stakeholders' participation are not included in the framework. In addition, two important aspects, namely dissemination of legitimate information among all stakeholders (information accessibility) and preventing the circulation of wrongful information (information misuse), are not considered in the strategy. In this study, the strategy was further reviewed by health care experts in the context of pandemic management and the need for integrating additional components was assessed using a questionnaire-based survey, which is discussed in the next section.

Figure 2. The National eHealth Strategy of Saudi Arabia. ICT: information and communications technology; MOH: Ministry of Health.



Methods

Survey

The purposes of this study were to review the current National eHealth Strategy of Saudi Arabia in the context of managing pandemics and to propose an integrated eHealth framework that can be used in the general health care context as well as for managing pandemics. The questionnaire was divided into four components. The first component focused on questions related to the current National eHealth Strategy (four items); the second component focused on the key considerations in an eHealth framework (nine items); the third component focused on key factors in the eHealth framework (one item); and the fourth component focused on assessing the need for an integrated eHealth framework for managing pandemics such as COVID-19 in Saudi Arabia (two items). The items in the questionnaire were developed based on the components reviewed in the background study and the National eHealth Strategy toolkit provided by the WHO [24]. A 5-point Likert scale [26] was used to collect the responses to the questions. The participants could present their opinions on five scales (1, strongly agree; 2, agree; 3, neutral; 4, disagree; 5, strongly disagree) relating to each item in the questionnaire. The questionnaire was translated into Arabic by a professional translator. However, both English and Arabic versions were used in the process of data collection. The survey was designed using the Google Surveys platform, and survey links to the English and Arabic versions were created to invite the participants. A pilot study was conducted with six health care professionals (three practitioners, two nurses, and one manager). Based on the feedback from the participants, a few changes in the questionnaire formulation and multiple-choice options were made to address grammatical errors. In addition, the Cronbach alpha for all the items in the four components was identified to be >0.85, revealing good consistency. The English version of the survey is provided in Multimedia Appendix 1.

Recruitment

Health care professionals were recruited using emails that included information about the purpose of the survey and the survey link. An additional note asking the participants to forward the email to their colleagues was included to increase the number of participants. Initially, using the portals at five hospitals (King Fahad Hospital in Jeddah, General Hospital in Medina, King Abdulaziz Medical City in Riyadh, King Khalid Ibn Abdul Aziz in HafarAlBatin, and King Khaled Hospital in Majma'ah), the email requesting participation in the survey was forwarded to all health care professionals working in these hospitals.

Sampling

This study used purposive sampling, whereby a nonprobability sample was obtained based on the purpose and the objective of the study, which mainly focused on reviewing the current National eHealth Strategy and proposing a new integrated eHealth framework for managing pandemics. Accordingly, the survey link was initially forwarded to 257 health care professionals working at five hospitals in Saudi Arabia. Snowball sampling is an effective technique to reach a larger sample population in a short time. In this technique, existing study subjects are requested to recruit future subjects from among their acquaintances [27]. As a result of using snowball sampling (requesting participants to forward emails to their colleagues in other hospitals), a final sample of 350 participants was obtained. Of these participants, 23 were removed due to low-quality responses and incomplete data, and 11 more were removed because they did not work for health care–related organizations. As a result, 316 eligible participants completed the survey.

Analytical Process

The survey was developed using the Google Surveys platform and was conducted from March 9 to April 6, 2020. Frequencies were calculated to analyze the collected data. The data were analyzed using four themes relevant to the study: review of the current National eHealth Strategy; objectives of the eHealth framework; importance of ICT; and need for an integrated eHealth framework in the context of managing pandemics in Saudi Arabia. The results are discussed in the following section.

Results

The final sample size recruited for this study was 316 respondents. Among the sample of respondents, 221/316 of the participants were male (69.9%), and 95/316 (30.1%) were female. In terms of age, the majority of the participants were aged 35-44 years (147/316, 46.5%), followed by 25-34 years (110/316, 34.8%), 45-54 years (30/316, 9.5%), and 18-24 years (15/316, 4.8); only 4.4% (14/316) of the participants were older than 54 years. In terms of education, the majority of the participants had the qualification of a bachelor's degree (161/316, 50.9%). Numerous participants held a master’s degree (74/316, 23.4%) or doctorate (24/316, 7.6%), reflecting the level of expertise of the participants. In addition, 18% (57/316) of the participants had a diploma or high school qualification. In terms of profession, the sample was evenly distributed across the relevant experts. Of the participants, 26.3% (83/316) worked in administrative departments, while 23.4% (74/316) were nurses, 20.6% (65/316) were physicians, and 19.9% (63/316) were technical experts. In addition, 9.8% (31/316) of the participants were involved in other health care activities, such as research and development, pharmaceutical operations, and academics. In terms of work experience, 54.8% (173/316) of the participants had more than 10 years of experience, while 27.5% (87/316) had 5-10 years of experience. This indicates that 82.2% (260/316) of the participants had more than five years of experience, reflecting that the majority of the participants were reliable. In addition, 12.3% (39/316) of the participants had less than two years of experience, and 5.4% (17/316) had two to five years of experience. Table 1 shows the frequency distributions of these variables.

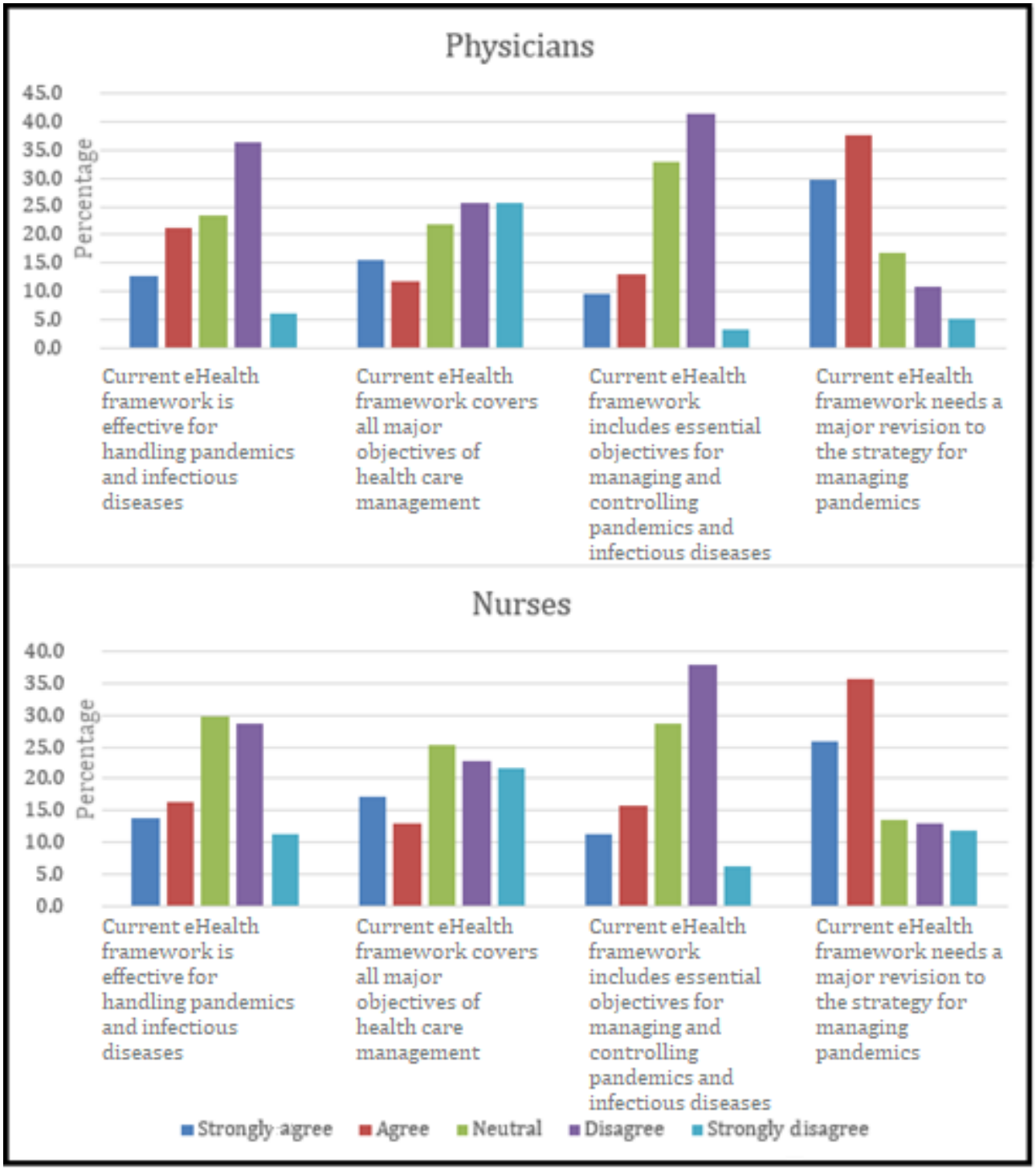
Table 1. Frequency distributions of key variables (N=316), n (%).

Variable	Value
Gender	
Male	221 (69.9)
Female	95 (30.1)
Age (years)	
18-24	15 (4.8)
25-34	110 (34.8)
35-44	147 (46.5)
45-54	30 (9.5)
>54	14 (4.4)
Education	
High school graduate or diploma	57 (18)
Bachelor's degree	161 (50.9)
Master's degree	74 (23.4)
Doctorate	24 (7.6)
Profession	
Physician	65 (20.6)
Nurse	74 (23.4)
Technical expert	63 (19.9)
Administrator	83 (26.3)
Other health care operation	31 (9.8)
Work experience (years)	
<2	39 (12.3)
2-5	17 (5.4)
5-10	87 (27.5)
>10	173 (54.8)

The next set of results focuses on reviewing the current national eHealth framework in Saudi Arabia. The effectiveness of the current framework in managing pandemics and infectious diseases was identified to be ineffective by the majority of the participants; 50.3% (159/316) of the participants disagreed and 8.9% (28/316) of the participants strongly disagreed with the statement that the current eHealth framework is effective, comprising nearly 60% of the total participants. Similarly, the statement that the current eHealth framework covers all major objectives of health care management received mixed responses. Although 40.2% (127/316) of the total participants disagreed and 7% (22/316) of the total participants strongly disagreed, a considerable number of participants (32%, 101/316) were neutral. However, the findings suggested that the current eHealth framework lacks a few objectives related to health care management. Similarly, considering the statement on inclusion of essential objectives for managing pandemics, 50% (158/316) of the participants disagreed, and 10.1% (22/316) of the participants strongly disagreed. Focusing on the need for revising the current framework, 53.8% (170/316) of the total participants strongly agreed and 16.5% (52/316) of the total participants agreed that the current framework should be revised, accounting for almost 70% of the total participants.

In addition, 91.4% (289/316) of the total participants agreed that there is need for a new effective and efficient eHealth framework for Saudi Arabia. It can be observed from Figure 3 that the opinions about the current eHealth framework slightly differed between physicians and nurses. A slightly greater number of physicians than nurses stated the opinion that the current eHealth framework is ineffective for managing pandemics. However, the majority of both physicians and nurses stated the opinion that the current eHealth framework needs major revisions.

Figure 3. Comparison of the opinions of physicians and nurses regarding the current eHealth framework in Saudi Arabia.



The next set of results (as shown in Table 2) focuses on the key considerations for developing a new integrated eHealth framework for managing pandemics in Saudi Arabia. In response to the statement that the new framework, in contrast to the existing framework, should focus on long-term objectives, 32.3% (101/316) of the participants strongly agreed with the statement, and 57.6% (182/316) of the participants agreed with it. Focusing on the distinction of services during normal and pandemic conditions may be useful to clearly outline the service delivery guidelines in these two conditions, which may clear any ambiguity regarding the operations and services. Accordingly, the idea of having different sets of objectives for the two conditions was supported by 94.6% (299/316) of the participants, among which 30.7% (97/316) strongly agreed with this statement and 63.9% (202/316) agreed with it. It is interesting to observe that 94% (297/316) of the total participants agreed that focusing on compatibility and integration is important. Accordingly, the need for inclusion of all stakeholders, including government, health care practitioners, businesses, and governments, and the need for clear roles and responsibilities of all stakeholders were supported by more than 85% of the total participants.

Table 2. Frequency distribution of key considerations related to the national eHealth framework of Saudi Arabia for managing pandemics (N=316), n (%).

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
The new framework should focus on long-term objectives	102 (32.3)	182 (57.6)	24 (7.6)	4 (1.3)	4 (1.3)
It should have different objectives relevant to managing health care services during pandemics	100 (31.7)	184 (58.2)	27 (8.5)	4 (1.3)	1 (0.3)
It should include regular health care objectives along with the objectives for controlling and managing pandemics	97 (30.7)	202 (63.9)	13 (4.1)	4 (1.3)	0 (0)
It should include all stakeholders: the public, health care practitioners, businesses, and government	109 (34.5)	166 (52.5)	33 (10.4)	7 (2.2)	1 (0.3)
It should specify clear roles and responsibilities for all stakeholders	121 (38.3)	175 (55.4)	17 (5.4)	0 (0)	3 (1.0)

The next set of results (as shown in Table 3) focuses on identifying the key factors to be considered in the framework for managing pandemics. The results demonstrate that public awareness, promotion of precautionary methods, dissemination of genuine information, increased stakeholder participation, increased reachability and accessibility of electronic services, improved communication, promotion of self-management and self-control approaches, formulation of new guidelines, and regular monitoring and updating of processes and approaches were identified to be the key factors in managing pandemics by the majority (>80%) of the total participants.

Table 3. Frequency distribution of key objectives to be included in the new eHealth framework (N=316), n (%).

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Develop awareness-raising programs	124 (39.2)	168 (53.2)	18 (5.7)	4 (1.3)	2 (0.6)
Promote precautionary methods during pandemics	136 (43.0)	168 (53.1)	8 (2.53)	2 (0.63)	2 (0.6)
Make accurate information accessible to the public	115 (36.4)	159 (50.3)	31 (9.8)	8 (2.5)	3 (1.0)
Increase participation of all stakeholders at individual, community, and national levels	115 (36.4)	173 (54.8)	19 (6.0)	4 (1.3)	5 (1.6)
Increase the richness and reachability of electronic services	127 (40.2)	166 (52.5)	14 (4.4)	6 (1.9)	3 (1.0)
Heighten consciousness and improve communication	140 (44.6)	176 (51.3)	11 (3.5)	0 (0)	2 (0.6)
Promote self-management and self-control approaches among individuals	98 (31.0)	176 (55.7)	31 (9.8)	7 (2.2)	4 (1.3)
Formulate new guidelines for managing health care services during pandemics	106 (33.5)	182 (57.6)	23 (7.3)	3 (1.0.)	2 (0.6)
Review the framework at regular intervals and update the objectives	106 (33.5)	182 (57.6)	24 (7.6)	1 (0.3)	3 (1.0)

Focusing on the use of ICT, it may be considered that ICT is effective in managing information flow and streamlining operations to increase efficiency. Accordingly, in relevance to the importance of ICT for tracking and monitoring the spread of infectious disease, the majority of the participants (74.1%, 234/316) considered it to be important. In view of the rising number of pandemics in the past few decades, the importance and immediate need for an eHealth framework for Saudi Arabia were assessed in the context of COVID-19. The findings revealed that more than three-fourths of the participants (78.8%, 249/316) agreed that developing an integrated framework is very important. The results of the survey reflect the need for an integrated eHealth framework and various components that need to be considered to manage pandemics. The findings are accordingly discussed in the next section.

Discussion

Principal Findings

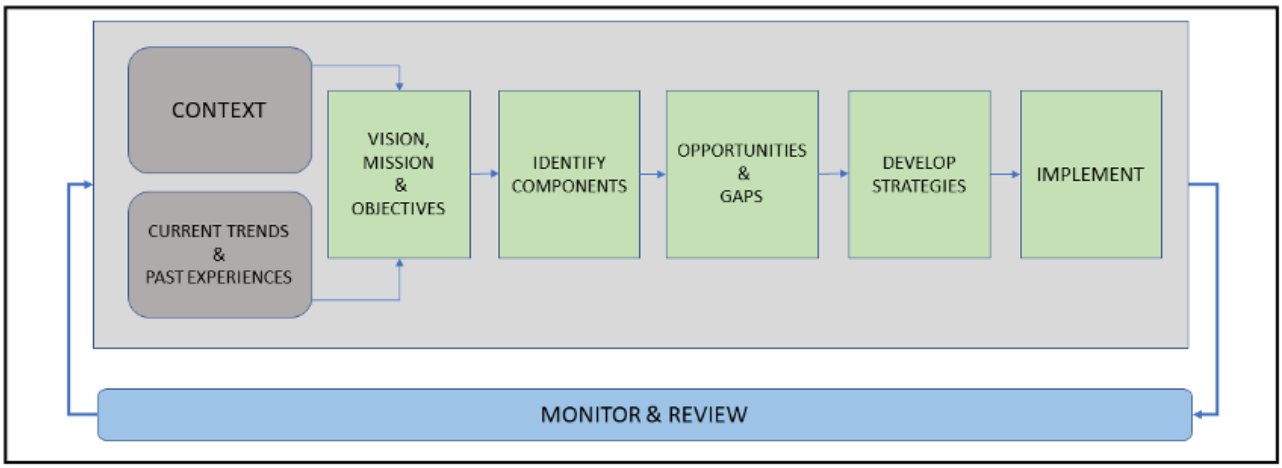
The results of this study clearly indicate that the current national eHealth strategy is ineffective for handling pandemics, as it does not incorporate all the necessary components and objectives for managing pandemics using eHealth strategies. In addition, 89.9% (284/316) of the survey participants were in favor of developing an integrated eHealth framework for managing pandemics. Focusing on these gaps, the necessary components of the framework were identified based on the findings (Table 2, Table 3) related to the key considerations and key factors to be included in the framework for managing and controlling pandemics. Firstly, it is important to determine the strategic context. The strategic context describes the priorities and challenges that eHealth can address. This context can be developed by reviewing the health statistics of the population, current health strategy, priorities, and goals. As identified from

the key considerations (Table 3), 88.9% (281/316) of the participants stated that long-term objectives should be considered. As pandemics can have long term impacts [28,29], the framework should focus on both short-term (as adopted in the current National eHealth Strategy) and long-term objectives. In addition, it is necessary to consider and review the current health care system. Accordingly, 93% (294/316) of the total participants stated that the objectives related to general health care and health care services during pandemics must be considered. In addition, the priorities and goals must be revised according to the situations that arise during pandemics. In addition, past trends and experiences in dealing with pandemics can be used as a valuable source of information for setting the context, priorities, and goals [29].

A vision statement presents a long-term objective, while a mission statement considers short-term objectives for achieving the vision. Managing and controlling epidemics requires short-term objectives that are effective in delivering health care services and preventing the spread of infectious disease using eHealth approaches. Based on the mission statement, the next step is to assess the components required to achieve the mission. These components can also be considered as building blocks for achieving the mission; they include leadership and governance, strategy and investment, services and applications, standards and interoperability, infrastructure, legislation, policy and compliance, and the workforce [24]. The leadership, standards and interoperability, investment, policy, and compliance components are ensured by governments and health care ministries when establishing the mission and overseeing all health care operations. Although infrastructure, services, and applications are related to the ICT environment, all other components can be considered as enablers in creating the environment for managing pandemics.

The next step is to analyze the opportunities, gaps, risks, and barriers for creating an eHealth environment for the components identified. While the opportunities suggest possible eHealth solutions and applications relevant to components, gaps can be identified in relation to services and infrastructure [29-32]. For instance, opportunities for collaboration between research and development and health care units such as hospitals, physicians, and pharmaceutical companies can be achieved by using a common electronic platform or portal for sharing information related to diseases, such as symptoms, diagnosis factors, effects of various medicines, and need for medicines and medical equipment. Lack of effective collaboration between these components can create risks and gaps related to the delivery of health care services; this can be assessed from the recent outbreak of COVID-19, where shortages of medical equipment and vaccines and a lack of information are clearly evident in various regions [32,33]. Based on the assessment of opportunities and gaps, strategies can be developed and implemented. As identified from the findings (Table 3), strategies such as promotion of precautionary methods, creating awareness, promotion of self-management and self-control practices, and increasing accessibility to genuine information were considered to be highly important during pandemics. One of the most important aspects in the process of developing strategies during pandemics is rapid monitoring and review of the overall development and implementation process. Accordingly, the whole approach is presented in Figure 4.

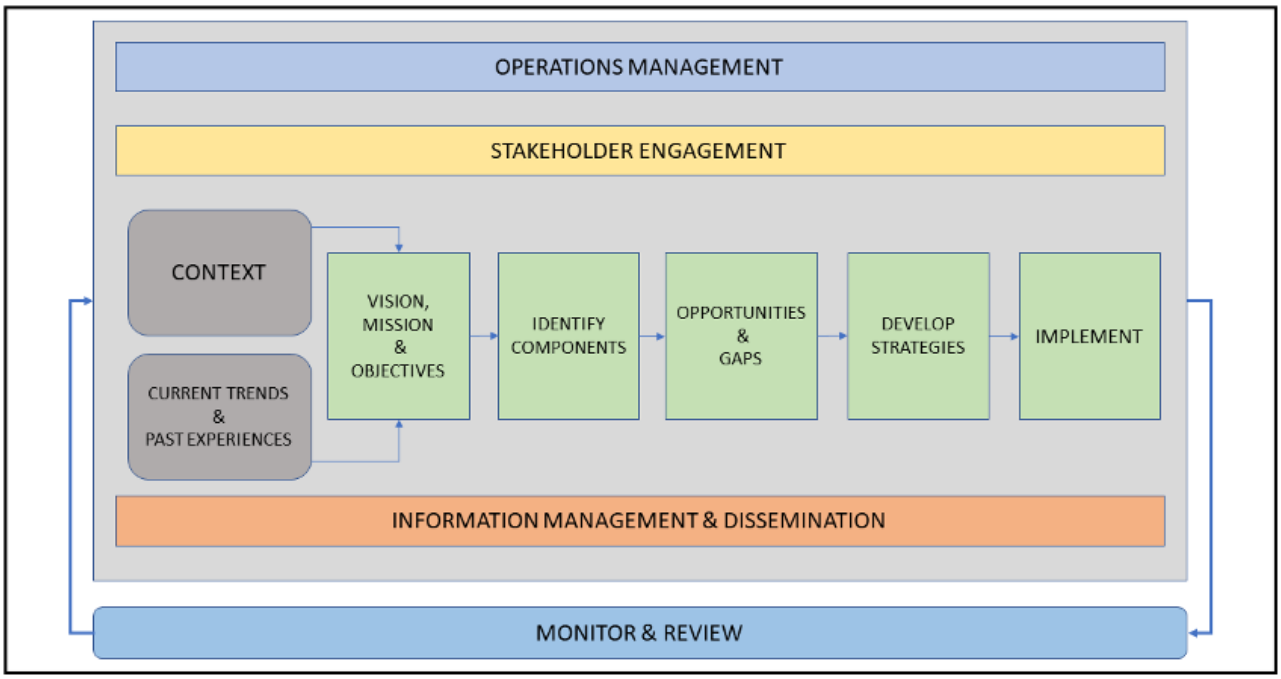
Figure 4. Approach for developing eHealth strategies during pandemics.



A few components must be managed throughout the process, including operations management, stakeholders' engagement, and information management, as shown in Figure 5, representing an integrated eHealth framework for managing and controlling pandemics in Saudi Arabia. Health care operations during pandemics involve service delivery and administration of health care activities; moreover, the scope of operations is increased by engaging all the stakeholders and their respective activities. This process requires high-level leadership and support, an appropriate governance structure, a multidisciplinary team, and an agreed timeline and resources for completing the tasks with maximum efficiency in the minimum amount of time [24]. For instance, the manufacturing of masks, testing kits, and medicines such as hydroxychloroquine is being implemented on a "wartime" footing in different countries for managing and controlling COVID-19 [34,35], and lack of effective leadership and strategy may result in various risks and challenges for controlling pandemics [36,37]. This represents a collaborative effort under the leadership of governments with various stakeholders, including physicians, pharmaceutical companies, and manufacturers of testing kits, operating with maximum

efficiency to address the needs of health care systems during a pandemic.

Figure 5. Proposed integrated eHealth framework for pandemic management.



Stakeholder engagement is another important component that must be managed during pandemics. The engagement process should focus on government's leadership roles, such as overseeing the engagement of all stakeholders; identifying different stakeholder groups, such as health care professionals, the public, pharmaceutical companies, medical equipment manufacturers, administrators, managers, and all other personnel involved with and related to health care activities; developing an approach for managing these groups; and defining the points of consultation and dissemination of information across all stakeholder groups. Some major stakeholder groups, as analyzed in [24], include decision-makers, key influencers, engaged stakeholders, broader stakeholders, and the general public. Leaders should ensure supportive and constructive engagement of all stakeholders during a pandemic, involving everyone in the process of managing and controlling the pandemic. Great benefits of ICT in health care management can be realized during pandemics. ICT can help prevent the spread of misinformation and can create awareness about infectious diseases, promote precautionary methods, provide accessibility of health care services through enhanced web-based or mobile applications, disseminate information and services in remote areas, and promote self-management and self-control practices during pandemics. All these factors were rated to be highly important by the majority of the survey participants (Table 3).

Thus, the three additional components are integrated with the strategy development components to form an integrated eHealth framework for managing and controlling pandemics. The process of review and monitoring is accordingly applied to all the components in the integrated eHealth framework, as shown in Figure 5. Accordingly, all the essential components of an eHealth strategy for managing pandemics are identified and integrated.

Limitations

In this study, various eHealth components required for managing and controlling pandemics were identified. However, functions within these components may differ from an application perspective in different regions. There is a need for clear definitions and explanations of these functions; these are not presented in this study, as it focused only on identifying the key components at the national level. In addition, the framework was developed specifically in a Saudi Arabian context based on a review of the current National eHealth Strategy; thus, it may not be applicable to other countries. However, it was ensured that the selection of components was generalized and focused on the necessity for health care management during pandemics.

Implications

Both theoretical and practical implications can be derived from this study. The literature review and the proposed framework can be used by researchers as a concept for developing and evaluating various strategies during pandemics, such as the lockdown strategy during the COVID-19 pandemic. In addition, the proposed framework offers valuable information for academicians and researchers regarding eHealth components and approaches for developing and implementing eHealth strategies. Practical implications include the consideration of the proposed framework by the Saudi Arabian government and the Ministry of Health in developing and implementing effective eHealth strategies during pandemics by relating the eHealth strategies to the various components proposed in the framework.

Future Research

The framework proposed in this paper is specific to Saudi Arabia. Future research will focus on evaluating the framework in the context of Saudi Arabia. Although the framework was

designed in the context of Saudi Arabia, it may be applicable to other regions with similar operational health care infrastructures and strategies. Therefore, future research may focus on validating the framework in similar countries in the Middle East and in other developing countries. In addition, both qualitative and quantitative methods can be used in the evaluation of the framework, which may lead to the collection of various types of data from which various inferences can be made. The author also proposes to extend the framework to various health care operations and the involvement of stakeholders (roles and responsibilities) using the approach of collective intelligence, where active engagement of all global stakeholders is considered to overcome the challenges of pandemics.

Conclusion

This study reviewed the current National eHealth strategy of Saudi Arabia; the review revealed various limitations and drawbacks in relation to health care management during epidemics. Accordingly, a survey instrument was administered to health care professionals to identify the necessary eHealth components and an approach was described for developing and implementing the strategies in relation to the identified components. The proposed framework is considered to be effective and efficient, as it is designed to be used for managing both general health care services and essential health care services during pandemics. This study proposes the main components of an eHealth framework for managing and controlling pandemics. Future work will focus on evaluating the identified components and identifying the necessary functions related to each component that are necessary during pandemics.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey questions.
[DOCX File , 314 KB - medinform_v8i11e19524_app1.docx]

References

1. Scott RE, Mars M. Principles and framework for eHealth strategy development. J Med Internet Res 2013 Jul 30;15(7):e155 [FREE Full text] [doi: 10.2196/jmir.2250] [Medline: 23900066]
2. Ouma S, Herselman M. E-health in Rural Areas: Case of Developing Countries. International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering 2008;2(4):304-310.
3. Wicks P, Stamford J, Grootenhuis MA, Haverman L, Ahmed S. Innovations in e-health. Qual Life Res 2014 Feb;23(1):195-203 [FREE Full text] [doi: 10.1007/s11136-013-0458-x] [Medline: 23852096]
4. Botha M, Botha A, Harselman M. The Benefits and Challenges of e-Health Applications: A Content Analysis of the South African context. In: Proceedings of the International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM 2014). 2014 Dec Presented at: International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM); December 12-14, 2014; Thessaloniki, Greece.
5. Pfeiffer KP, Giest S, Dumortier J, Artmann J. Country Brief: Austria. eHealth Strategies. 2010 Apr 09. URL: http://ehealth-strategies.eu/database/documents/Austria_CountryBrief_eHStrategies.pdf [accessed 2020-09-04]
6. Country Reports Database: Individual eHealth Strategies Country Reports and further information. eHealth Strategies. URL: http://ehealth-strategies.eu/database/database.html [accessed 2020-04-09]
7. Li J, Ray P, Seale H, MacIntyre R. An E-Health Readiness Assessment Framework for Public Health Services—Pandemic Perspective. In: Proceedings of the 45th Hawaii International Conference on System Sciences. 2012 Presented at: 45th Hawaii International Conference on System Sciences; January 4-7, 2012; Maui, HI p. 2800-2809. [doi: 10.1109/hicss.2012.95]
8. Li J, Land LPW, Ray P, Chattopadhyaya S. E-Health readiness framework from Electronic Health Records perspective. IJIEM 2010;6(4):326. [doi: 10.1504/ijiem.2010.035626]
9. Fengou M, Mantas G, Lymberopoulos D, Komninos N, Fengos S, Lazarou N. A New Framework Architecture for Next Generation e-Health Services. IEEE J Biomed Health Inform 2013 Jan;17(1):9-18. [doi: 10.1109/titb.2012.2224876]
10. Li J, Ray P, Bakshi A, Seale H, MacIntyre R. Tool for E-Health Preparedness Assessment in the Context of an Influenza Pandemic. IJEHMC 2013;4(2):18-33. [doi: 10.4018/jehmc.2013040102]
11. Kiberu VM, Mars M, Scott RE. Development of an evidence-based e-health readiness assessment framework for Uganda. Health Inf Manag 2019 Apr 22:1833358319839253. [doi: 10.1177/1833358319839253] [Medline: 31010314]
12. National E-Health Strategy. Saudi Arabia Ministry of Health. URL: https://www.moh.gov.sa/en/Ministry/nehs/Pages/How-we-developed-the-eHealth-Strategy.aspx [accessed 2020-04-10]
13. Lipsitch M. Seasonality of SARS-CoV-2: Will COVID-19 go away on its own in warmer weather? Center for Communicable Disease Dynamics. URL: https://ccdd.hsph.harvard.edu/will-covid-19-go-away-on-its-own-in-warmer-weather/ [accessed 2020-04-10]
14. eHealth at WHO. World Health Organization. URL: https://www.who.int/ehealth/about/en/ [accessed 2020-04-11]

15. Golby J. Lockdown's hottest viral trends: raging at the neighbours and torching 5G towers. The Guardian. 2020 Apr 06. URL: https://www.theguardian.com/commentisfree/2020/apr/06/lockdown-viral-trends-5g-towers-coronavirus-cases [accessed 2020-04-11]
16. Thoughts on healthcare management in an epidemic. Deloitte. URL: https://www2.deloitte.com/cn/en/pages/risk/articles/thoughts-on-construction-of-large-health-management-system-under-2019-ncov.html [accessed 2020-04-10]
17. Hospital Preparedness for Epidemics. World Health Organization. 2014 Apr 04. URL: https://apps.who.int/iris/rest/bitstreams/674837/retrieve [accessed 2020-10-04]
18. Mental health and COVID-19. World Health Organization. URL: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/technical-guidance/mental-health-and-covid-19 [accessed 2020-07-15]
19. Asmundson GJ, Taylor S. How health anxiety influences responses to viral outbreaks like COVID-19: What all decision-makers, health authorities, and health care professionals need to know. J Anxiety Disord 2020 Apr;71:102211 [FREE Full text] [doi: 10.1016/j.janxdis.2020.102211] [Medline: 32179380]
20. Hii A, Chughtai AA, Housen T, Saketa S, Kunasekaran MP, Sulaiman F, et al. Epidemic intelligence needs of stakeholders in the Asia-Pacific region. Western Pac Surveill Response J 2018 Dec 18;9(4):28-36 [FREE Full text] [doi: 10.5365/wpsar.2018.9.2.009] [Medline: 30766745]
21. Dasaklis TK, Pappis CP, Rachaniotis NP. Epidemics control and logistics operations: A review. Int J Prod Econ 2012 Oct;139(2):393-410. [doi: 10.1016/j.ijpe.2012.05.023]
22. Büyüktahtakın İE, des-Bordes E, Kıbış EY. A new epidemics–logistics model: Insights into controlling the Ebola virus disease in West Africa. Eur J Oper Res 2018 Mar;265(3):1046-1063. [doi: 10.1016/j.ejor.2017.08.037]
23. 600 people die in Iran from drinking neat alcohol to cure coronavirus. Middle East Monitor. 2020 Apr 08. URL: https://www.middleeastmonitor.com/20200408-600-people-die-in-iran-from-drinking-neat-alcohol-to-cure-coronavirus/ [accessed 2020-04-09]
24. National eHealth Strategy Toolkit. World Health Organization. 2012. URL: https://apps.who.int/iris/bitstream/handle/10665/75211/9789241548465_eng.pdf;jsessionid=7EE3B7E95FBD9B273845987853441982?sequence=1 [accessed 2020-04-03]
25. Lahey T, Nelson W. A Dashboard to Improve the Alignment of Healthcare Organization Decisionmaking to Core Values and Mission Statement. Camb Q Healthc Ethics 2019 Dec 20;29(1):156-162. [doi: 10.1017/s0963180119000884]
26. Likert R. A Technique for the Measurement of Attitudes. Arch Psychology 1932;22(140):55.
27. Goodman LA. Snowball Sampling. Ann Math Statist 1961 Mar;32(1):148-170. [doi: 10.1214/aoms/1177705148]
28. Guimbeau A, Menon N, Musacchio A. The Brazilian Bombshell? The Long-Term Impact of the 1918 Influenza Pandemic the South American Way. NBER Working Papers. 2020 Apr. URL: https://www.nber.org/system/files/working_papers/w26929/w26929.pdf [accessed 2020-10-26]
29. Pecchia L, Pallikarakis N, Magjarevic R, Iadanza E. Health Technology Assessment and Biomedical Engineering: Global trends, gaps and opportunities. Med Eng Phys 2019 Oct;72:19-26 [FREE Full text] [doi: 10.1016/j.medengphy.2019.08.008] [Medline: 31554572]
30. Belflower R, Burt B, Burdsall DP, Cullen D, Hill E, Sanku G, et al. Identifying Opportunities for Targeted Interventions: Gaps in Endocavity Probe High-Level Disinfection Practices Across Healthcare Settings in Illinois. Am J Infect Control 2018 Jun;46(6):S9. [doi: 10.1016/j.ajic.2018.04.021]
31. Pennathur PR, Herwaldt LA. Role of Human Factors Engineering in Infection Prevention: Gaps and Opportunities. Curr Treat Options Infect Dis 2017 May 6;9(2):230-249 [FREE Full text] [doi: 10.1007/s40506-017-0123-y] [Medline: 32226329]
32. Vanderpuye V, Elhassan MMA, Simonds H. Preparedness for COVID-19 in the oncology community in Africa. Lancet Oncol 2020 May;21(5):621-622. [doi: 10.1016/s1470-2045(20)30220-5]
33. Sendolo JM. CSO Platform Frowns at 'Poor' Information Management in COVID-19 Fight. Daily Observer. 2020 Mar 26. URL: https://www.liberianobserver.com/news/cso-platform-frowns-at-poor-information-management-in-covid-19-fight/ [accessed 2020-04-13]
34. Alter C. 'It's Mass Confusion.' Small Manufacturers Look for Leadership as They Make Medical Supplies to Fight Coronavirus. Time. 2020 Mar 24. URL: https://time.com/5808500/small-manufacturers-coronavirus/ [accessed 2020-04-14]
35. Hufford A. New Manufacturers Jump Into Mask Making as Coronavirus Spreads. Wall Street Journal. 2020 Mar 21. URL: https://www.wsj.com/articles/new-manufacturers-jump-into-mask-making-as-coronavirus-spreads-11584792003 [accessed 2020-04-14]
36. Baker P, Rogers K, Enrich D, Haberman M. Trump's Aggressive Advocacy of Malaria Drug for Treating Coronavirus Divides Medical Community. The New York Times. 2020 Apr 06. URL: https://www.nytimes.com/2020/04/06/us/politics/coronavirus-trump-malaria-drug.html [accessed 2020-04-13]
37. Selinger H. Stealing masks and stockpiling hydroxychloroquine — what America has become during this epidemic is deeply worrying. The Independent. 2020 Apr 07. URL: https://www.independent.co.uk/voices/coronavirus-us-masks-trump-hydroxychloroquine-covid-19-drug-a9450261.html [accessed 2020-04-13]

Abbreviations

ICT: information and communications technology

MERS: Middle East respiratory syndrome
SARS: severe acute respiratory syndrome
WHO: World Health Organization

Edited by G Eysenbach; submitted 21.04.20; peer-reviewed by I Ur Rehman, D Sobnath; comments to author 13.07.20; revised version received 16.07.20; accepted 14.09.20; published 26.11.20.

Please cite as:
Alsharif A
Applying eHealth for Pandemic Management in Saudi Arabia in the Context of COVID-19: Survey Study and Framework Proposal
JMIR Med Inform 2020;8(11):e19524
URL: http://medinform.jmir.org/2020/11/e19524/
doi:10.2196/19524
PMID:33035174

Original Paper

Analysis of the Trends in Publications on Clinical Cancer Research in Mainland China from the Surveillance, Epidemiology, and End Results (SEER) Database: Bibliometric Study

Min-Qiang Lin[1*], MSc; Chen-Lu Lian[2*], MD; Ping Zhou[2*], MD; Jian Lei[3], MD; Jun Wang[2], MD; Li Hua[3], MD; Juan Zhou[3], MD; San-Gang Wu[2], MD

[1]Department of Scientific Management, The First Affiliated Hospital of Xiamen University, Xiamen, China
[2]Department of Radiation Oncology, The First Affiliated Hospital of Xiamen University, Xiamen, China
[3]Department of Obstetrics and Gynecology, The First Affiliated Hospital of Xiamen University, Xiamen, China
[*]these authors contributed equally

Corresponding Author:
San-Gang Wu, MD
Department of Radiation Oncology, The First Affiliated Hospital of Xiamen University
55 Zhenhai Road
Xiamen
China
Phone: 86 5922139531
Email: wusg@xmu.edu.cn

Abstract

Background: The application of China's big data sector in cancer research is just the beginning. In recent decades, more and more Chinese scholars have used the Surveillance, Epidemiology, and End Results (SEER) database for clinical cancer research. A comprehensive bibliometric study is required to analyze the tendency of Chinese scholars to utilize the SEER database for clinical cancer research and provide a reference for the future of big data analytics.

Objective: Our study aimed to assess the trend of publications on clinical cancer research in mainland China from the SEER database.

Methods: We performed a PubMed search to identify papers published with data from the SEER database in mainland China until August 31, 2020.

Results: A total of 1566 papers utilizing the SEER database that were authored by investigators in mainland China were identified. Over the past years, significant growth in studies based on the SEER database was observed (P<.001). The top 5 research topics were breast cancer (213/1566, 13.6%), followed by colorectal cancer (185/1566, 11.8%), lung cancer (179/1566, 11.4%), gastrointestinal cancer (excluding colorectal cancer; 149/1566, 9.5%), and genital system cancer (93/1566, 5.9%). Approximately 75.2% (1178/1566) of papers were published from the eastern coastal region of China, and Fudan University Shanghai Cancer Center (Shanghai, China) was the most active organization. Overall, 267 journals were analyzed in this study, of which Oncotarget was the most contributing journal (136/267, 50.9%). Of the 1566 papers studied, 585 (37.4%) were published in the second quartile, 489 (31.2%) in the third quartile, 312 (19.9%) in the first quartile, and 80 (5.1%) in the fourth quartile, with 100 (6.4%) having an unknown Journal Citation Reports ranking.

Conclusions: Clinical cancer research based on the SEER database in mainland China underwent constant and rapid growth during recent years. High-quality and comprehensive cancer databases based on Chinese demographic data are urgently needed.

(JMIR Med Inform 2020;8(11):e21931) doi:10.2196/21931

KEYWORDS

cancer; China; data collection; bibliometrics; PubMed; SEER program

Introduction

Background

The incidence of human cancer is increasing worldwide. It is estimated that the global burden of cancer will increase by more than 60% by 2040 [1]. Cancer has been an important public health problem in low- and middle-income countries, as well as in upper-middle-income countries. Cancer research has become one of the leading research fields of bioscience around the world, and the number of publications on cancer increases at a rate of more than 2% per year [2].

A high-quality cancer database can provide researchers with convenient data analysis and build a sharing platform among researchers, which could pave the way for revealing the mechanism of tumorigenesis and its progression [3]. To share clinical data with different regions, some other countries plunged into building multicenter databases far ahead of China [4]. In 1973, the US National Cancer Institute combined the tumor registration stations in several regions to form the Surveillance, Epidemiology, and End Results (SEER) database. The SEER program is a globally accessible authoritative cancer database representing approximately 34.6% of the US population, which includes non-Hispanic White, non-Hispanic Black, Hispanic, and Asian populations [5]. The SEER program collects data on patient demographics, tumor location, tumor stage, first course of therapy, and vital status. The SEER database is a valuable, population-based resource that can be used to study the diagnosis and treatment across demographic characteristics and geographic areas, and it has become a unique research resource for oncology practice. It provides morbidity and mortality data on various histopathological subtypes, and data on molecular characteristics are also expanding. The database is being further developed to capture other biomarker data and the results of specific populations, and to expand the biobank to support cutting-edge cancer research that can improve oncology practices. Therefore, the SEER program plays an important role in clinical cancer research, public health management, and policy making [5].

In recent years, China has made significant progress in clinical cancer research, and many studies have gained international recognition [6-10]. Although major hospitals have established databases in China, they have not shared their research findings with one another. Since the SEER database is a globally accessible authoritative cancer database, more and more scholars, particularly those from China, have used it to conduct clinical cancer research in recent years. There are a lot of differences between the United States and China, including their population's genetic makeup, health system, health services, health insurance, health policies, socioeconomic status, and culture. Therefore, the research findings in the SEER cancer registries may not be generalizable to the people of China. However, several recent studies found that the characteristics of the data from the SEER program were consistent with those from Chinese institutions [11-14]. There is currently no comprehensive bibliometric study that has characterized the clinical cancer research in China based on the SEER database. Carrying out a comprehensive bibliometric analysis is helpful to analyze the contribution of Chinese scholars to clinical cancer research and provide a specific clinical reference for the future of big data analytics.

Objective

This study aimed to evaluate the characteristics of clinical cancer research using SEER data from mainland China using a bibliometric approach.

Methods

Search Strategy

Using the search terms "Surveillance, Epidemiology, and End Results or SEER, and China," we identified related publications from the PubMed database before August 31, 2020. PubMed is a free, publicly available database established by the US National Library of Medicine [15]. As one of the largest databases in the life science and biomedical fields worldwide, it comprises more than 30 million biomedical abstracts from journals and online books.

Publications with first authors or corresponding authors affiliated with mainland Chinese institutions were included in the study, while those that utilized the SEER-Medicare database (SEER-Medicare data are not released outside of the United States) and special types of publications, including comments, letters, and reviews, were excluded. Papers from Taiwan Province, Hong Kong, and Macao Special Administrative Regions were also excluded from this analysis. This study was approved by the ethics committee of the First Affiliated Hospital of Xiamen University (Xiamen, Fujian, China).

Indices

The indices analyzed in this study included the number and trend of publications, research topics, type of affiliation (university, hospital, or other research center), geographical distribution, journal, Journal Citation Reports (JCR) ranking, and status of international cooperation.

Statistical Analysis

Descriptive statistical analyses were used in this study. Characteristics were evaluated and analyzed using Microsoft Excel, and linear regression was performed using SPSS statistical software (version 22.0; IBM Corp). P values <.05 were considered statistically significant.

Results

Number and Trend of Publications

The flowchart of publication selection for this study is shown in Figure 1. We retrieved a total of 1667 publications, of which 1566 publications were included in this study. Figure 2 shows the trend in publications on clinical cancer research using the SEER database in mainland China (R^2=0.430, P<.001). Chinese authors first used the SEER database in 1999. The single paper in the SEER database that was published in 1999 was authored by Guo et al [16], who were affiliated with the People's Hospital of Beijing Medical University and had collaborated with Huvos and colleagues from Memorial Sloan-Kettering Cancer Center

in the United States. Interestingly, there were no papers from Chinese institutions that utilized the SEER database during the 11 years from 2000 to the end of 2011. However, the number of papers published per year increased rapidly from 2012 to 2019: 2012 (n=6), 2013 (n=12), 2014 (n=19), 2015 (n=43), 2016 (n=96), 2017 (n=181), 2018 (n=288), and 2019 (n=459). Despite the COVID-19 pandemic in 2020, there were 461 publications as of August 31, 2020.

Figure 1. Flowchart of publication selection for the study.

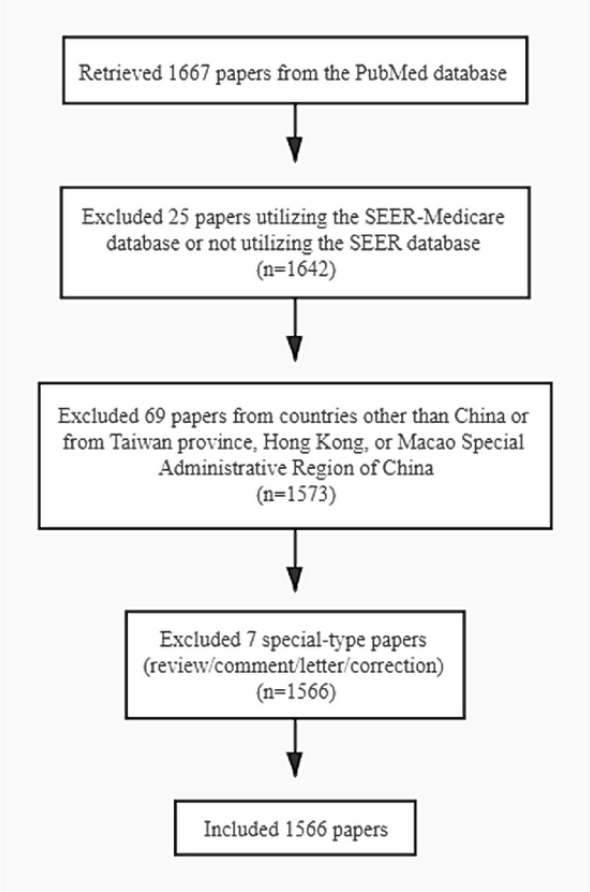
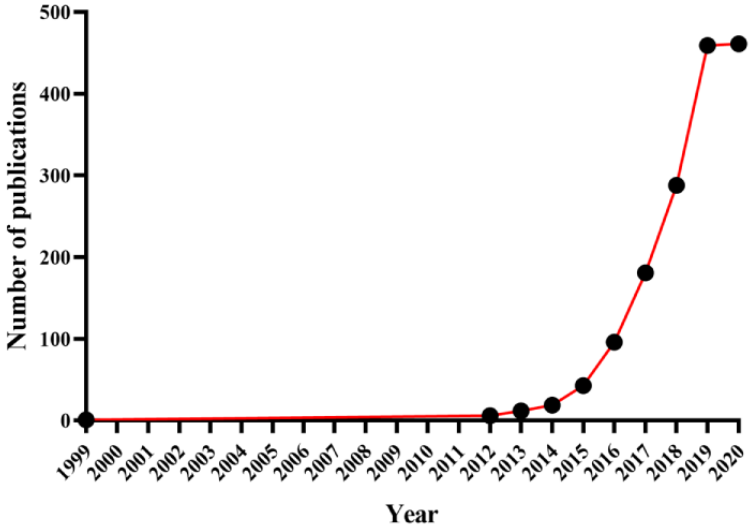


Figure 2. The trends in publications on clinical cancer research in mainland China from the Surveillance, Epidemiology, and End Results database between 1999 and 2020.



Research Topics

Table 1 presents the main research topics of the 1566 papers included in the study. Breast cancer was the most frequently researched topic (213/1566, 13.6%), followed by colorectal cancer (185/1566, 11.8%), lung cancer (179/1566, 11.4%), gastrointestinal cancer (excluding colorectal cancer; 149/1566, 9.5%), and genital system cancer (93/1566, 5.9%). These top

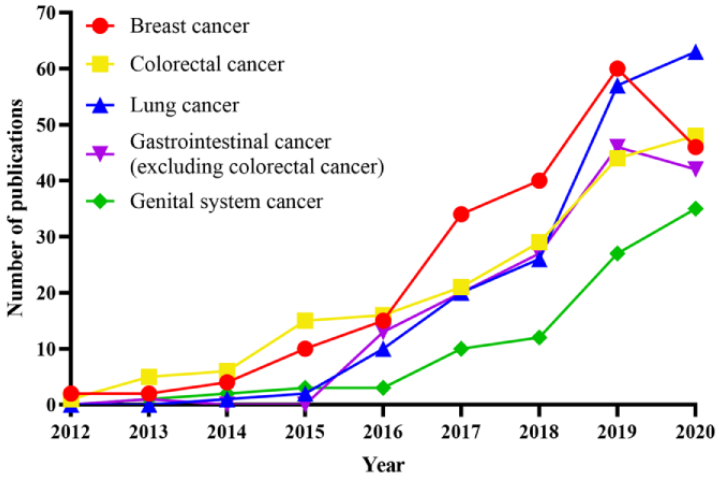
5 research topics were the focus of 52.3% (819/1566) of the included papers.

Figure 3 shows the publication trends on the top 5 cancer sites in mainland China over time. It shows an increasing interest in research exploring breast cancer, lung cancer, gastrointestinal cancer (excluding colorectal cancer), colorectal cancer, and genital system cancer (in both genders). Chinese researchers were also studying rare cancers. A total of 164 papers discussed rare cancers, such as pulmonary lymphoepithelioma-like carcinoma, spindle cell carcinoma, thymoma, and adenoid cystic carcinoma of the breast, and this number was expanding rapidly in recent years.

Table 1. Main cancer research topics of publications in mainland China identified in the Surveillance, Epidemiology, and End Results database.

Main cancer research topics	Number of publications
Breast cancer	213
Colorectal cancer	185
Lung cancer	179
Gastrointestinal cancer (excluding colorectal cancer)	149
Genital system cancer	93
Pancreatic cancer	83
Liver cancer	72
Thyroid cancer	62
Esophageal cancer	62
Bone cancer	49

Figure 3. The trends in publications on the top 5 cancer sites in mainland China from the Surveillance, Epidemiology, and End Results database between 2012 and 2020.

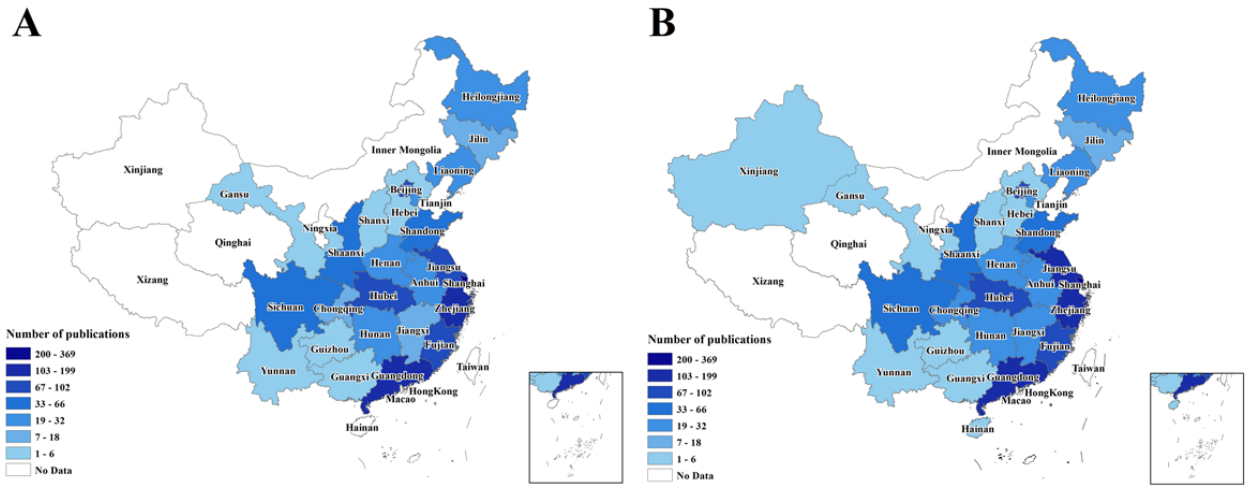


Geographical Distribution of Publications

Figure 4A shows the distribution of the publications using the SEER database whose first authors were affiliated with Chinese institutions. These first authors who used the SEER database to publish papers came from 25 provinces and municipalities of China. Among all of these regions, first authors were most often affiliated with institutions in Shanghai (369/1566, 23.6%), followed by Guangdong (199/1566, 12.7%), Zhejiang (174/1566, 11.1%), Hubei (102/1566, 6.5%), and Jiangsu (102/1566, 6.5%).

Figure 4B shows the distribution of publications using the SEER database whose corresponding authors were affiliated with Chinese institutions. These corresponding authors who used the SEER database to publish papers came from 27 provinces or municipalities of China. A similar distribution was observed in the papers whose corresponding authors were affiliated with Chinese institutions, mostly concentrated in Shanghai, Guangdong, Zhejiang, Jiangsu, and Hubei.

Figure 4. The distribution of the publications stratified by first authors (A) and corresponding authors (B) using the Surveillance, Epidemiology, and End Results database.



Top 10 Most Contributing Author Affiliations

The top 10 affiliated organizations of the most contributing first authors and corresponding authors extracted from the 1566 publications are presented in Tables 2 and 3. More than 75% (1178/1566) of papers were published from the eastern coastal region of China, Overall, the organization that was affiliated with the top contributing first authors was Fudan University Shanghai Cancer Center (163/1566, 10.4%), followed by Sun Yat-sen University Cancer Center (73/1566, 4.7%), the First Affiliated Hospital of Xi'an Jiaotong University (48/1566, 3.1%), West China Hospital of Sichuan University (42/1566, 2.7%), the Second Affiliated Hospital of Zhejiang University School of Medicine (41/1566, 2.6%), and the First Affiliated Hospital of Xiamen University (41/1566, 2.6%). All affiliations of first authors extracted from the 1566 publications (excluding three papers whose first authors were affiliated with the United States) were classified into three sectors (university, hospital, and government agency). The most contributing institution sector was hospital (1501/1563, 96.0%), followed by university (60/1563, 3.8%), and government agency (2/1563, 0.1%). Of the 1501 papers from hospitals, 1459 (97.2%) publications were produced from the best tertiary hospitals, and 1475 (98.3%) were from hospitals affiliated with universities.

The organization that was affiliated with the top contributing corresponding authors was Fudan University Shanghai Cancer Center (165/1566, 10.5%), followed by Sun Yat-sen University Cancer Center (77/1566, 4.9%), Zhongshan Hospital of Fudan University (43/1566, 2.7%), the First Affiliated Hospital of Xi'an Jiaotong University (43/1566, 2.7%), and the Second Affiliated Hospital of Zhejiang University School of Medicine (42/1566, 2.7%). All affiliations of corresponding authors extracted from the 1566 papers (excluding 35 publications' organization affiliations of corresponding authors affiliated with the United States, Germany, Australia, and Japan) were also classified into three sectors (university, hospital, and government agency). Similarly, the most contributing institution sector was hospital (1474/1531, 96.3%), followed by university (54/1531, 3.5%), and government agency (3/1531, 0.2%). In the 1474 papers whose corresponding authors came from hospitals, 1440 (97.7%) publications were from best tertiary hospitals, and 1458 (98.9%) were from hospitals affiliated with universities.

Table 2. Top 10 affiliated organizations of the most contributing first authors.

Affiliations of first authors	Number of publications
Fudan University Shanghai Cancer Center	163
Sun Yat-sen University Cancer Center	73
The First Affiliated Hospital of Xi'an Jiaotong University	48
West China Hospital of Sichuan University	42
The Second Affiliated Hospital of Zhejiang University	41
The First Affiliated Hospital of Xiamen University	41
Zhongshan Hospital of Fudan University	39
Affiliated Union Hospital of Fujian Medical University	33
Union Hospital of Tongji Medical College of Huazhong University of Science and Technology	32
Renmin Hospital of Wuhan University	27

Table 3. Top 10 affiliated organizations of the most contributing corresponding authors.

Affiliations of corresponding authors	Number of publications
Fudan University Shanghai Cancer Center	165
Sun Yat-sen University Cancer Center	77
The First Affiliated Hospital of Xi'an Jiaotong University	43
Zhongshan Hospital of Fudan University	43
The Second Affiliated Hospital of Zhejiang University	42
West China Hospital of Sichuan University	42
Union Hospital of Tongji Medical College of Huazhong University of Science and Technology	33
Affiliated Union Hospital of Fujian Medical University	31
Renmin Hospital of Wuhan University	27
Shandong Cancer Hospital Affiliated to Shandong University	27

Journals and Journal Visibility

Overall, 267 journals were analyzed in the present study, of which the most contributing journal was Oncotarget (136/267, 50.9%), followed by Cancer Medicine (98/267, 36.7%), Journal of Cancer (76/267, 28.5%), Medicine (Baltimore) (72/267, 27.0%), and Cancer Management and Research (69/267, 25.8%). The JCR ranking was retrieved from Web of Science, a global citation database. The top 30 most common journals extracted from the 1566 publications are shown in Figure 5. The distribution of the JCR ranking of the 1566 publications is depicted in Figure 6. A total of 37.4% (585/1566) of the publications were published in the second quartile, followed by the third quartile (489/1566, 31.2%), first quartile (312/1566, 19.9%), unknown JCR ranking (100/1566, 6.4%), and fourth quartile (80/1566, 5.1%).

Figure 5. The distribution of the publications stratified by the top 30 most common journals.

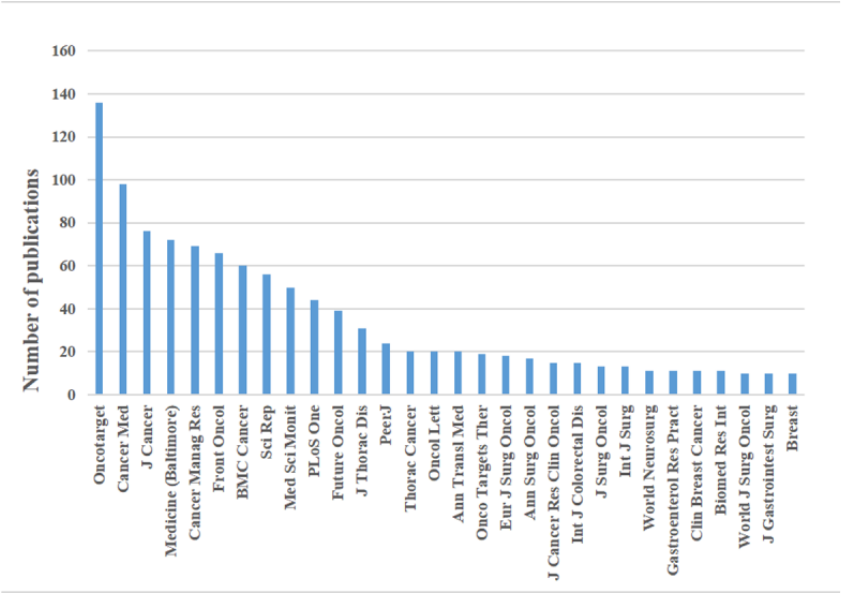
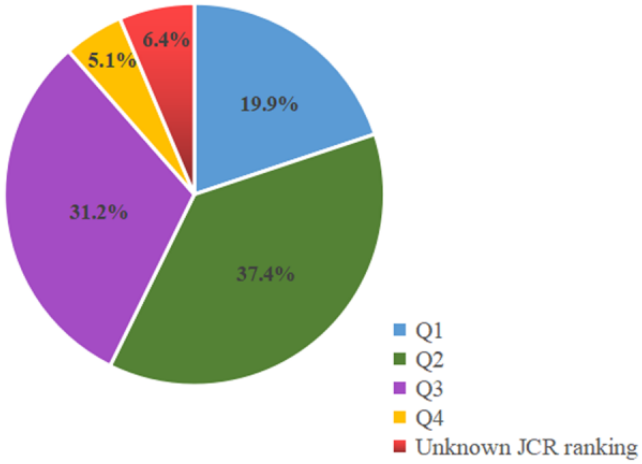


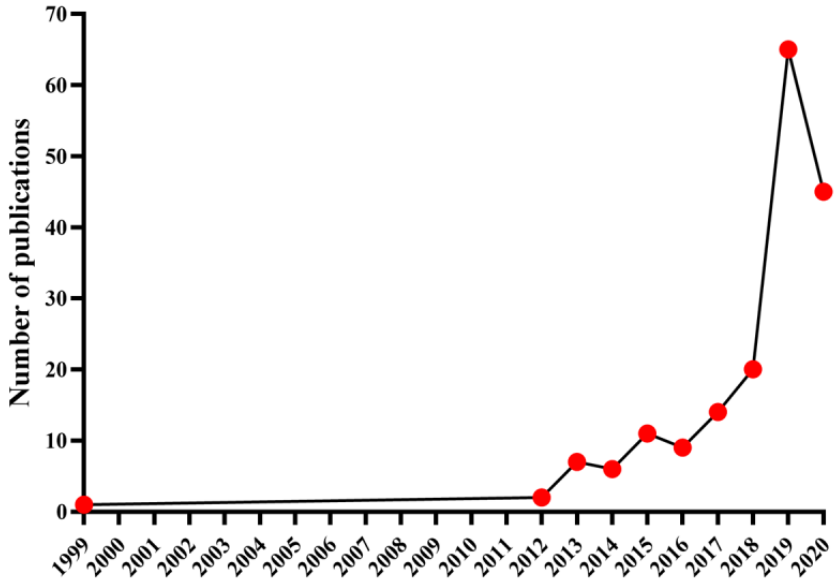
Figure 6. The distribution of the 1566 papers stratified by journal rankings published from Journal Citation Reports (JCR). Q1: first quartile; Q2: second quartile; Q3: third quartile; Q4: fourth quartile.



International Cooperation

A total of 180 publications were jointly produced by organizations from China and 23 other countries including the United States (114/180, 63.3%), Russia (17/180, 9.4%), Italy (12/180, 6.7%), Australia (10/180, 5.6%), and Japan (10/180, 5.6%). Chinese institutions first cooperated with institutions from other countries using the SEER database in 1999. Similarly, no papers were jointly produced by institutions from China and other countries utilizing the SEER database during the 11 years from 2000 to the end of 2011. However, the number of papers produced through international cooperation increased substantially from 2016 (9/1566, 0.6%) to 2019 (65/1566, 4.2%). As of August 31, 2020, a total of 45 papers had been published in the year 2020 with international cooperation (Figure 7).

Figure 7. The trends in publications with international cooperation using the Surveillance, Epidemiology, and End Results database between 1999 and 2020.



Discussion

The large sample size of the SEER database and strong statistical guarantees for current cancer concerns add much clinical value to research based on the SEER database. However, the application of China's big data sector in cancer research is just the beginning. In the recent decade, more and more Chinese scholars have used the SEER database for clinical cancer research. A comprehensive bibliometric study is required to analyze the tendency of Chinese scholars to utilize the SEER database for clinical cancer research and provide a reference for the future of big data analytics. In the present study, we completed a bibliometric study of the SEER database-related publications in China, some prominent characteristics of which were found. First, SEER database-related publications have made continuous and rapid growth in China in recent years. Second, the research topics mainly focused on high-incidence and high-mortality cancers, such as breast cancer, colorectal

cancer, lung cancer, gastrointestinal cancer (excluding colorectal cancer), and genital system cancer. Third, publications regarding clinical cancer research using the data from the SEER program were mainly from the best tertiary hospitals in the eastern coastal region of China.

Our study revealed a rapidly growing interest in clinical cancer research from the SEER program in China. With the exception of 1 paper published in 1999, Chinese scholars did not utilize the SEER database for clinical cancer research until 2012. This trend has grown rapidly since 2012, reaching 459 papers in the year 2019. The Covid-19 pandemic has changed almost every aspect of life and society in the year 2020. However, 461 publications were published in the first 8 months of 2020, indicating that the COVID-19 pandemic has not had a negative impact on publications in the SEER database. This phenomenon might attribute to the development of cancer research and informationization in China. Over the past decades, Chinese researchers have gradually made more in-depth use of cancer databases and noticed the important role of the SEER program. This finding was consistent with other studies that have demonstrated an increase in cancer research papers from China [17,18]. In addition, we noticed that not only Chinese scholars, but also scholars from France [19,20], South Korea [21,22], Japan [23], Italy [24], and Switzerland [25], have used the SEER database to conduct clinical cancer research.

With respect to the research topics, the SEER database-related publications were unbalanced, mainly concentrated on breast cancer, colorectal cancer, lung cancer, gastrointestinal cancer, and genital system cancer. Publications on these five cancer sites of the SEER database accounted for 52.3% of all 1566 publications. This regular pattern of distribution of the research topics was consistent with the top 5 cancer sites for estimated cases worldwide for both sexes, which were lung cancer, breast cancer, colorectal cancer, prostate cancer, and stomach cancer [1]. Although these top 5 research topics have been widely studied, Chinese researchers' interest in them has increased steadily. Moreover, the findings from the SEER studies have been introduced into the treatment guidelines of the National Comprehensive Cancer Network (NCCN) Guidelines and the European Society for Medical Oncology (ESMO) Guidelines, which demonstrate the critical role of the SEER database in clinical cancer research [26,27].

In addition to playing a significant role in the study of these high-incidence cancers, the SEER database is often used to explore rare diseases. Because rare cancers are very uncommon, it is difficult to collect data from a single institution. The SEER database provided substantial valuable data for the study of rare cancers and avoided the selection bias found in single-center retrospective studies. Nevertheless, the results from the SEER database are also retrospective, and prospective studies are needed to validate the results from the SEER database. However, rare cancers are challenging to study prospectively because of their low incidence. The latest NCCN and ESMO guidelines cited several publications on rare cancers based on SEER data, such as male breast cancer, occult breast cancer, and mesenchymal chondrosarcoma [28-30]. Thus, these publications based on SEER data provided insight to help clarify the characteristics, treatment protocols, prognostic indicators, and risk stratification of rare cancers [31,32].

With respect to geographical and institutional distribution, papers published from China using the SEER database were extremely unbalanced. More than 75% of papers were published from the eastern coastal region of China, which was the area with the highest incidence of cancer morbidity in mainland China [33]. In our analysis, more than 90% of publications were produced by the best tertiary hospitals, while 4 of the top 5 contributing organizations were located in the eastern coastal region of China. Regarding the population and economic levels, high-quality medical resources are distributed unevenly in China, with 71 of the 2018 top 100 hospitals in China situated in the eastern coastal region of China [34-37]. In addition, most medical colleges in China are located in this region and could provide resources for cancer research [38]. By relying on universities, researchers at affiliated hospitals can get more research support.

The findings from the SEER program may contribute to cancer prevention and treatment, whether the cancer is common or rare. Nevertheless, the SEER program still has its limitations in that its vital statistics only consist of death and survival data, and there are no data regarding locoregional recurrence and distant metastasis after adopting the corresponding treatment. In addition, the SEER database only represents people who live in the United States. Whether the results of SEER-based research can be applied to other countries, especially to areas of high-incidence cancers like China, needs to be verified by data from other countries. Some Chinese researchers combined the SEER database with their databases to conduct studies and found that the characteristics of the data from the SEER program were consistent with that from Chinese institutions [11-14]. This indicated that the SEER database could provide valuable data for clinical oncology treatment in China. The above results indicate that clinical cancer research based on the SEER database may provide a valuable reference for clinical cancer researchers in China.

China should establish cancer databases based on its demographic characteristics. To our knowledge, Chinese cancer registration work was started much later than in other upper middle-income countries [4]. The number of cancer registries in China has increased rapidly since 2002, while a significant gap still exists between China and other upper middle-income countries. At present, the items and contents of population-based cancer data in China are limited. Therefore, high-quality cancer databases based on Chinese demographic data are urgently needed to better reflect the oncology practices in China [39-42].

The limitation of our analysis is that we only collected publications from the PubMed database, which may ignore some publications that were not indexed in PubMed, resulting in incomplete data.

Conclusions

In conclusion, our study suggests that clinical cancer research regarding the SEER database has rapidly increased in China in the past decade. High-quality and national comprehensive cancer

registries from China are needed to provide a reference for the future of big data analytics.

Authors' Contributions

MQL, CLL, PZ, and SGW were responsible for the study design; MQL, CLL, PZ, and SGW contributed to writing the report; and JW, JL, LH, and JZ contributed to data analysis. All authors have reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424 [FREE Full text] [doi: 10.3322/caac.21492] [Medline: 30207593]
2. Lewison G, Purushotham A, Mason M, McVie G, Sullivan R. Understanding the impact of public policy on cancer research: a bibliometric approach. Eur J Cancer 2010 Mar;46(5):912-919. [doi: 10.1016/j.ejca.2009.12.020] [Medline: 20064708]
3. Yang J, Cai HY. [The Cancer-related Bioinformatics Databases] [In Chinese]. Biotechnology Bulletin 2015 Nov;31(11):89-101. [doi: 10.13560/j.cnki.biotech.bull.1985.2015.11.010]
4. Yang L. [Brief Review on Cancer Registration at Home and Abroad] [In Chinese]. China Cancer 2005 Dec;14(12):772-775. [doi: 10.3969/j.issn.1004-0242.2005.12.001]
5. The Surveillance, Epidemiology, and End Results (SEER) Program Overview. National Cancer Institute. URL: https://seer.cancer.gov/about/overview.html [accessed 2019-12-20] [WebCite Cache ID https://seer.cancer.gov/about/overview.html]
6. Zhang Y, Chen L, Hu GQ, Zhang N, Zhu XD, Yang KY, et al. Gemcitabine and Cisplatin Induction Chemotherapy in Nasopharyngeal Carcinoma. N Engl J Med 2019 Sep 19;381(12):1124-1135. [doi: 10.1056/NEJMoa1905287] [Medline: 31150573]
7. Chen Y, Ye J, Zhu Z, Zhao W, Zhou J, Wu C, et al. Comparing Paclitaxel Plus Fluorouracil Versus Cisplatin Plus Fluorouracil in Chemoradiotherapy for Locally Advanced Esophageal Squamous Cell Cancer: A Randomized, Multicenter, Phase III Clinical Trial. J Clin Oncol 2019 Jul 10;37(20):1695-1703 [FREE Full text] [doi: 10.1200/JCO.18.02122] [Medline: 30920880]
8. Ma F, Ouyang Q, Li W, Jiang Z, Tong Z, Liu Y, et al. Pyrotinib or Lapatinib Combined With Capecitabine in HER2-Positive Metastatic Breast Cancer With Prior Taxanes, Anthracyclines, and/or Trastuzumab: A Randomized, Phase II Study. J Clin Oncol 2019 Oct 10;37(29):2610-2619. [doi: 10.1200/JCO.19.00108] [Medline: 31430226]
9. Luo Y, Luo L, Wampfler JA, Wang Y, Liu D, Chen Y, et al. 5-year overall survival in patients with lung cancer eligible or ineligible for screening according to US Preventive Services Task Force criteria: a prospective, observational cohort study. Lancet Oncol 2019 Aug;20(8):1098-1108 [FREE Full text] [doi: 10.1016/S1470-2045(19)30329-8] [Medline: 31255490]
10. Li Y, Wu YL. The second wave of checkpoint inhibitors with chemotherapy for advanced non-small-cell lung cancer. Lancet Oncol 2019 Jul;20(7):889-891. [doi: 10.1016/S1470-2045(19)30148-2] [Medline: 31122902]
11. Hua J, Zhang B, Xu J, Liu J, Ni Q, He J, et al. Determining the optimal number of examined lymph nodes for accurate staging of pancreatic cancer: An analysis using the nodal staging score model. Eur J Surg Oncol 2019 Jun;45(6):1069-1076 [FREE Full text] [doi: 10.1016/j.ejso.2019.01.018] [Medline: 30685327]
12. Shi X, Liu XK, An CM, Wei WJ, Tao Y, Ji Y, et al. Anatomic extent of lymph node metastases as an independent prognosticator in node-positive major salivary gland carcinoma: A study of the US SEER database and a Chinese multicenter cohort. Eur J Surg Oncol 2019 Nov;45(11):2143-2150. [doi: 10.1016/j.ejso.2019.06.029] [Medline: 31253544]
13. He J, Tsang JY, Xu X, Li J, Li M, Chao X, et al. AJCC 8th edition prognostic staging provides no better discriminatory ability in prognosis than anatomical staging in triple negative breast cancer. BMC Cancer 2020 Jan 06;20(1):18 [FREE Full text] [doi: 10.1186/s12885-019-6494-3] [Medline: 31906874]
14. Liang W, He J, Shen Y, Shen J, He Q, Zhang J, et al. Impact of Examined Lymph Node Count on Precise Staging and Long-Term Survival of Resected Non-Small-Cell Lung Cancer: A Population Study of the US SEER Database and a Chinese Multi-Institutional Registry. J Clin Oncol 2017 Apr 10;35(11):1162-1170 [FREE Full text] [doi: 10.1200/JCO.2016.67.5140] [Medline: 28029318]
15. PubMed. URL: https://www.ncbi.nlm.nih.gov/pubmed/ [accessed 2019-12-20]
16. Guo W, Xu W, Huvos AG, Healey JH, Feng C. Comparative frequency of bone sarcomas among different racial groups. Chin Med J (Engl) 1999 Dec;112(12):1101-1104. [Medline: 11721448]
17. Zhang LP. [Oncology literature in China from 1994 to 2003: a bibliometric analysis] [In Chinese]. Chin J Med Libr Inf Sci 2006 Jan;15(1):67-69.

18. Zheng JJ, Zhang HR, Jing S. [A bibliometric analysis of oncology papers published by Chinese authors covered in SCI from 2010 to 2012] [In Chinese]. Chin J Med Libr Inf Sci 2012 Dec;21(12):64-71. [doi: 10.3969/j.issn.1671-3982.2012.12.022]
19. Benoit L, Pauly L, Phelippeau J, Koskas M. Impact of Sociodemographic Characteristics on the Quality of Care in the Surgical Management of Endometrial Cancer: An Analysis of a National Database in the United States. Gynecol Obstet Invest 2020 Mar;85(3):222-228. [doi: 10.1159/000506048] [Medline: 32224609]
20. Gonthier C, Douhnai D, Koskas M. Lymph node metastasis probability in young patients eligible for conservative management of endometrial cancer. Gynecol Oncol 2020 Apr;157(1):131-135. [doi: 10.1016/j.ygyno.2020.02.021] [Medline: 32139150]
21. Kim BH, Kim S, Kim YI, Chang JH, Hwang K, Kim S, et al. Development of an Individualized Prediction Calculator for the Benefit of Postoperative Radiotherapy in Patients with Surgically Resected De Novo Stage IV Breast Cancer. Cancers (Basel) 2020 Jul 29;12(8):2103 [FREE Full text] [doi: 10.3390/cancers12082103] [Medline: 32751136]
22. Jung J, Kim BH, Kim J, Oh S, Kim SJ, Lim CS, et al. Validating the ACOSOG Z0011 Trial Result: A Population-Based Study Using the SEER Database. Cancers (Basel) 2020 Apr 11;12(4):950 [FREE Full text] [doi: 10.3390/cancers12040950] [Medline: 32290437]
23. Watanuki R, Hayashida T, Yokoe T, Kawai Y, Kikuchi M, Nakashoji A, et al. Impact of neoadjuvant and adjuvant chemotherapy on invasive lobular carcinoma: A propensity score-matched analysis of SEER data. Breast J 2020 May 25 (forthcoming). [doi: 10.1111/tbj.13884] [Medline: 32449173]
24. Rosiello G, Palumbo C, Pecoraro A, Luzzago S, Deuker M, Stolzenbach LF, et al. The effect of sex on disease stage and survival after radical cystectomy: a population-based analysis. Urol Oncol 2020 Oct 06 (forthcoming). [doi: 10.1016/j.urolonc.2020.09.004] [Medline: 33036900]
25. Siebenhüner AR, Güller U, Warschkow R. Population-based SEER analysis of survival in colorectal cancer patients with or without resection of lung and liver metastases. BMC Cancer 2020 Mar;20(1):246. [doi: 10.1186/s12885-020-6710-1] [Medline: 32293337]
26. Petkov VI, Miller DP, Howlader N, Gliner N, Howe W, Schussler N, et al. Breast-cancer-specific mortality in patients treated based on the 21-gene assay: a SEER population-based study. NPJ Breast Cancer 2016 Jun;2:16017. [doi: 10.1038/npjbcancer.2016.17] [Medline: 28721379]
27. Qiu M, Hu J, Yang D, Cosgrove DP, Xu R. Pattern of distant metastases in colorectal cancer: a SEER based study. Oncotarget 2015 Nov;6(36):38658-38666. [doi: 10.18632/oncotarget.6130] [Medline: 26484417]
28. Cloyd JM, Hernandez-Boussard T, Wapnir IL. Outcomes of partial mastectomy in male breast cancer patients: analysis of SEER, 1983-2009. Ann Surg Oncol 2013 May;20(5):1545-1550. [doi: 10.1245/s10434-013-2918-5] [Medline: 23460016]
29. Wu SG, Zhang WW, Sun JY, Li FY, Lin HX, Chen YX, et al. Comparable Survival between Additional Radiotherapy and Local Surgery in Occult Breast Cancer after Axillary Lymph Node Dissection: A Population-based Analysis. J Cancer 2017 Oct;8(18):3849-3855. [doi: 10.7150/jca.21217] [Medline: 29151972]
30. Schneiderman BA, Kliethermes SA, Nystrom LM. Survival in Mesenchymal Chondrosarcoma Varies Based on Age and Tumor Location: A Survival Analysis of the SEER Database. Clin Orthop Relat Res 2017 Mar;475(3):799-805 [FREE Full text] [doi: 10.1007/s11999-016-4779-2] [Medline: 26975384]
31. Zhuo M, Zheng Q, Chi Y, Jia B, Zhao J, Wu M, et al. Survival analysis via nomogram of surgical patients with malignant pleural mesothelioma in the Surveillance, Epidemiology, and End Results database. Thorac Cancer 2019 May;10(5):1193-1202 [FREE Full text] [doi: 10.1111/1759-7714.13063] [Medline: 30951250]
32. Yin Z, Wang Y, Wu Y, Zhang X, Wang F, Wang P, et al. Age distribution and age-related outcomes of olfactory neuroblastoma: a population-based analysis. Cancer Manag Res 2018 May;10:1359-1364 [FREE Full text] [doi: 10.2147/CMAR.S151945] [Medline: 29881306]
33. Zheng RS, Sun KX, Zhang SW, Zeng HM, Zou XN, Chen R, et al. [Report of cancer epidemiology in China, 2015] [In Chinese]. Zhonghua Zhong Liu Za Zhi 2019 Jan 23;41(1):19-28. [doi: 10.3760/cma.j.issn.0253-3766.2019.01.005] [Medline: 30678413]
34. Yue LR, Sun LQ, Yang K, Li Z. [Research on the Status Quo and Strategy of China's Medical and Health System Reform under the Background of Reform and Opening-up] [In Chinese]. China Health Industry 2019 Jul;16(19):185-188. [doi: 10.16659/j.cnki.1672-5654.2019.19.185]
35. Shi BG, Wu SL. [Spatial Analysis of Access to Care:Geographic Distribution of China's Top Hospitals from The Perspective of Embedded Stratification Theory] [In Chinese]. Journal of Gansu Administration Institute 2019 Oct;5:94-104.
36. An YF. [Characters and Improvement Strategies of Distribution of High-Quality Medical Resources] [In Chinese]. Chinese Health Quality Management 2011 Sep;18(5):110-113. [doi: 10.13912/j.cnki.chqm.2011.05.034]
37. Zhuang YQ, Liao XB, Wang XL, Yao SF. [Hospital Blue Book: China Hospital Competitiveness Report (2018-2019)] [In Chinese]. Beijing: Social Sciences Academic Press (China); 2019.
38. ShanghaiRanking. URL: https://www.shanghairanking.com.cn/ [accessed 2019-12-20]
39. Wei WQ, He J. [Some thoughts on cancer registry in China: in the era of big data and informatization] [In Chinese]. Zhonghua Zhong Liu Za Zhi 2019 Jan 23;41(1):15-18. [doi: 10.3760/cma.j.issn.0253-3766.2019.01.004] [Medline: 30678412]
40. Zhang SW, Chen WQ, Wang L. [The 30 Years of Cancer Registration in China] [In Chinese]. China Cancer 2009 Apr;18(4):256-259. [doi: 10.3969/j.issn.1004-0242.2007.07.001]

41. Chen W. [Establishing and Perfecting of Cancer Registration System in China] [In Chinese]. China Cancer 2011 Jan;20(1):7-9.
42. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA Cancer J Clin 2016 Mar;66(2):115-132 [FREE Full text] [doi: 10.3322/caac.21338] [Medline: 26808342]

Abbreviations

ESMO: European Society for Medical Oncology
JCR: Journal Citation Reports
NCCN: National Comprehensive Cancer Network
SEER: Surveillance, Epidemiology, and End Results

Edited by C Lovis; submitted 29.06.20; peer-reviewed by P Yang, S Song; comments to author 06.09.20; revised version received 19.10.20; accepted 25.10.20; published 17.11.20.

Please cite as:
Lin MQ, Lian CL, Zhou P, Lei J, Wang J, Hua L, Zhou J, Wu SG
Analysis of the Trends in Publications on Clinical Cancer Research in Mainland China from the Surveillance, Epidemiology, and End Results (SEER) Database: Bibliometric Study
JMIR Med Inform 2020;8(11):e21931
URL: http://medinform.jmir.org/2020/11/e21931/
doi:10.2196/21931
PMID:33200992

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org